EVALUATING THE EFFECTIVENESS OF AN ONLINE STANDARDIZATION PLATFORM FOR ENGLISH AS A FOREIGN LANGUAGE (EFL) ORAL EXAMINERS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

GÖKHAN YILDIZ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER EDUCATION AND INSTRUCTIONAL TECHNOLOGY

AUGUST 2019

Approval of the thesis:

EVALUATING THE EFFECTIVENESS OF AN ONLINE STANDARDIZATION PLATFORM FOR ENGLISH AS A FOREIGN LANGUAGE (EFL) ORAL EXAMINERS

submitted by **GÖKHAN YILDIZ** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Education and Instructional Technology Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Ömer Delialioğlu
Head of Department, **Comp. Edu. and Inst. Tech.**

Assist. Prof. Dr. Cengiz Savaş Aşkun
Supervisor, **Comp. Edu. and Inst. Tech., METU**

Examining Committee Members:

Assist. Prof. Dr. Özlem Canaran
DFL, University of Turkish Aeronautical Association

Assist. Prof. Dr. Cengiz Savaş Aşkun
Comp. Edu. and Inst. Tech., METU

Assist. Prof. Dr. Göknur Kaplan
Comp. Edu. and Inst. Tech., METU

Date: 02.08.2019

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Gökhan Yıldız

Signature:

ABSTRACT

EVALUATING THE EFFECTIVENESS OF AN ONLINE STANDARDIZATION PLATFORM FOR ENGLISH AS A FOREIGN LANGUAGE (EFL) ORAL EXAMINERS

Yıldız, Gökhan
Master of Science, Computer Education and Instructional Technology
Supervisor: Assist. Prof. Dr. Cengiz Savaş Aşkun

August 2019, 163 pages

This study aimed to evaluate the effectiveness of an online standardization platform prior to speaking exams at a foundation university in Ankara. The study was conducted in December 2018 and twenty-four instructors of English participated in the study. This study investigated how consistent instructors were in face-to-face and online standardization groups in grading ten sample oral exams within their groups. Intraclass Correlation Coefficient (ICC) analysis was conducted for both groups. The results demonstrated that online and face-to-face standardization groups scored ten sample exams consistently with excellent agreement values (ICC > .90). This study also aimed to find out whether there was a significant difference between the scores of oral exams given by the raters trained in a face-to-face standardization training and the ones trained on an online standardization platform. The results of the study showed that there was no significant difference considering the abovementioned condition. Another objective of the study was to see whether there was an agreement (consistency) between instructors in the official oral examination of the school. The Kappa statistics suggested that the majority of pairs in the official examination scored students consistently with a substantial agreement value (K > .60). In addition, semi-structured interviews with the instructors from both mediums revealed that instructors appreciated the content of the training to a great extent. Instructors in the online standardization platform reported their positive feelings about the platform such as instructional design, layout and practicality of the platform, and the flexibility of time and place.

Keywords: Online Standardization, Standardizing Speaking Exams, Teacher Professional Development, Inter-rater Reliability (Agreement), Testing Speaking

ÖZ

YABANCI DİL OLARAK İNGİLİZCE (EFL) KONUŞMA SINAVLARI DEĞERLENDİRİCİLERİ İÇİN HAZIRLANAN ÇEVRİMİÇİ BİR STANDARDİZASYON PLATFORMUNUN ETKİNLİĞİNİN DEĞERLENDİRİLMESİ

Yıldız, Gökhan
Yüksek Lisans, Bilgisayar ve Öğretim Teknolojileri Eğitimi
Tez Danışmanı: Dr. Öğr. Üyesi Cengiz Savaş Aşkun

Ağustos 2019, 163 sayfa

Bu çalışma, Ankara'daki bir vakıf üniversitesinde konuşma sınavlarından önce yapılan web tabanlı bir standardizasyon platformunun etkinliğini araştırmayı amaçlamaktadır. Bu araştırma Aralık 2018 tarihinde yapılmış olup araştırmaya 24 İngilizce öğretim görevlisi katılmıştır. Bu çalışma, yüz yüze ve online standardizasyon eğitimi alan öğretim görevlilerinin kendi gruplarındaki on örnek sözlü sınavı puanlandırmada ne kadar tutarlı olduklarını araştırmıştır. Her iki grup için Sınıf İçi Korelasyon Katsayısı (ICC) analizi yapılmıştır. Sonuçlar, çevrimiçi ve yüz yüze standardizasyon gruplarının, mükemmel uyum değerleri ile tutarlı bir şekilde on örnek sınavı puanlandırdıklarını göstermiştir (ICC > .90). Bu çalışma aynı zamanda yüz yüze standardizasyon eğitimi almış öğretmenlerin ve çevrimiçi standardizasyon platformunda eğitilen öğretim görevlilerinin puanlandırdıkları sözlü sınav puanları arasında anlamlı bir fark olup olmadığını tespit etmeyi amaçlamaktadır. Çalışmanın sonuçları, yukarıda belirtilen durum göz önüne alındığında önemli bir fark olmadığını göstermiştir. Çalışmanın diğer bir amacı da, okulun resmi sözlü sınavında öğretim görevlileri arasında bir anlaşma (tutarlılık) olup olmadığını görmekti. Kappa istatistikleri, resmi sınavdaki çiftlerin çoğunun, öğrencileri iyi bir anlaşma katsayısıyla tutarlı bir şekilde puanlandırdığını göstermiştir (K > .60). Ayrıca, her iki standardizasyon türünden gelen öğretim görevlileriyle yapılan yarı-yapılandırılmış görüşmeler, öğretim görevlilerinin eğitimin içeriğini büyük ölçüde takdir ettiğini ortaya koydu. Çevrimiçi standardizasyon platformundaki öğretim görevlileri, platformun öğretim tasarımı, düzeni ve uygulanabilirliği ile zaman ve yer kolaylığı gibi platform hakkındaki olumlu düşüncelerini bildirmişlerdir.

Anahtar Kelimeler: Web-tabanlı Standardizasyon, Konuşma Sınavlarının Standardizasyonu, Öğretmen Profesyonel Gelişim, Değerlendiriciler Arası Güvenilirlik (Anlaşma), İngilizce Konuşma Sınavlar

To my mother, my sister and my beloved cat Müjgan

ACKNOWLEDGEMENTS

Throughout the process of completing this study, there have been many stressful moments. I would very much like to thank my supervisor, Assist. Prof. Dr. Cengiz Savaş Aşkun, for his contributions and support to deal with those moments and finish this process with ease. He always welcomed all my questions and showed great patience to me. I feel lucky and privileged to have a chance to work with him throughout this study.

I would also like to thank my committee member Assist. Prof. Dr. Göknur Kaplan for her precious contributions and feedbacks for my study, and her friendly manner when I approached to talk to her and sparing her valuable time to chat with me.

I also wish to extend my sincere gratitude to my manager and Thesis Examining Committee member, Assist. Prof. Dr. Özlem Canaran, for her valuable and support during the process of writing my thesis. If it had not been her, it would not have been possible for me to write this thesis study.

Finally, I would like to thank to my cat, Müjgan, for putting up with me while I was writing this study. Had she not ignored me while I was studying, it would not have been quite possible for me to complete this study

.

TABLE OF CONTENTS

APPENDICES

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

ABBREVIATIONS

DE: Distance Education

PD: Professional Development

TPD: Teacher Professional Development

OS: Online Standardization

FFS: Face-to-face Standardization

LIST OF SYMBOLS

SYMBOLS

< : Less than

> : More than

≤ : Equal or less than

CHAPTER 1

INTRODUCTION

This chapter provides information about the background of the study, statement of the problem, the purpose of the study, the significance of the study and the research questions.

1.1. Background of the Study

Semmar (2013) defined adult education as the attainment of knowledge or sets of skills realized through receiving formal education in educational institutions, enrolling in activities targeting adults, and educating oneself by means of self-study. The latter option, which involves self-instruction, has been increasingly opted by adult learners over the last decades, and its utilization through distance education modalities has been becoming prevalent throughout the world. According to Bernard et al. (2009), web-based learning, a type of distance education (DE), has become a popular option of education as it enabled learners to communicate with the instructors although it was exercised lacking physical presence of tutors. Furthermore, what made DE a mainstream educational option in educational fields was the use of online educational mediums together with higher internet speed compared to past. As a result, the utilization of web-based learning courses started to emerge in many different fields such as online high schools, university courses, and continuous professional development (Peters, 2003).

While there have been various types of DE in the past such as correspondence letters sent to learners via postal services (Bernard et al., 2009), asynchronous DE where learners are educated through text-based materials, pre-recorded videos, and educational software, which enables learners to receive education one-way lacking the communicative options with instructors (Holden & Westfall, 2010), and lastly synchronous DE where learners are given the opportunity communicate with instructors through interaction, collaboration, and feedback (Peterson & Bond, 2004) by means of video-conferencing, online discussion platforms, electronic mails, notes, electronic chats, and presentations, all of which are provided to learners online (Skylar et al., 2017).

The synchronous or web-based type of DE has been favored more by learners and educators as it enabled a learner-centered approach where learners are able to satisfy their educational needs through online communities, flexible study scheduling, up-to-date and more developed implementations (Boisselle, 2014). As a result of these opportunities, online training was also seen as a way to implement job training for people in the workplace in order to provide skills and information needed to keep up with the trends and developments in various industries apart from educational institutions (Clegg, Hudson, & Steel, 2003).

Edinger (2017) emphasized that online training provides teachers with an opportunity to receive information about the fundamentals of their work, job requirements, and enhancing their teaching skills without dependence on conventional face-to-face trainings. It is a widely accepted notion that the quality of teaching could be boosted by professional development (PD) in various institutions (Kennedy, 2016; Penuel, Fishman, Yamaguchi, & Gallagher, 2007; Sun, Penuel, Frank, Gallagher, & Youngs, 2013). Improving instructor quality is considered as an impressively powerful way to boost students' academic performances. Therefore, consistent PD programs are needed to enhance educators' knowledge, their classroom practices and their performance on testing administrations (Masters, De Kramer, O'Dwyer, Dash, & Russell, 2010).

Badri, Alnuaimi, Mohaidat, Yang, and Al Rashedi (2016) also suggested that having a quality supportive environment needed by teachers could be achieved by creating and developing productive PD programs. It was also mentioned that PD activities could enhance teachers’ skills and expertise in their fields since they are required to keep up with constantly changing teaching practices and student profiles.

Since the integration of technology into the realm of education and PD, there has been a rapid increase in the number of web-based training platforms aiming to enhance pre-service and in-service teachers’ knowledge about their fields and instructional practices. Along with the developments in technology, web and computer-based platforms have proven to be an effective means of delivery to people thanks to their feasibility and ease of accessibility (Mixon, 2017).

In the study by Means, Tomaya, Murphy, Bakia, and Jones (2009), web-based training platforms were described as a bridge between standardization in curricular operations and teachers as they enable institutions to disseminate materials and objectives in a fast and quality way. Furthermore, standardizing curricular operations such as assessments plays a vital role in educational institutions since it might dramatically affect the validity of tests and perceptions about how fair these tests are. Especially in productive tests such as written and oral assessments, teachers are appointed as raters to grade test taker’s performances. Therefore, quality of grading is highly crucial for schools to ensure the validity and reliability of exam criteria and results (Ling, Mollaun, & Xi, 2014; Yan, 2014). In addition, Yan (2014) discussed that scores of productive tests are able to show differences between the raters. Hence, it might affect the overall score of test takers decreasing the reliability of tests. Therefore, teachers as raters of these productive assessment types should be trained so efficiently that they would not face difficulties while they are scoring test takers’ performances and nor should their particular scores vary regarding the overall quality of these exams.

As mentioned before, teachers can be trained through various PD activities both in person and online. So, it is up to institutions whether to conduct a PD activity that requires all teachers to show up in person or one that teachers can access via the Internet (Kunnavatana, Bloom, Samaha, & Dayton, 2013).

1.2. Statement of the Problem

The application of quality PD programs such as teacher training requires well-planning and suitable time allocations so that teachers can participate when they do not have lessons. However, there are some challenges which hinder the utilization of such programs. Institutions with a high number of instructors, for example, may find it quite challenging to gather all teachers to be trained in one specific place as these instructors' schedules might be busy and there might be clashes between their class hours and the PD session. The study (Masters et al., 2010) supported these ideas as institutions might be affected negatively when PD programs are conducted at a time that will force teachers to leave their classrooms.

Moreover, PD programs aiming to standardize scoring of productive tests such as writing and speaking exams would require all teachers to come together in a place (Gradel & Edson, 2012). One disadvantage of this might be the noise problem as trainees should share ideas and information between the trainer and their peers (Keis, Grab, Schneider, & Öchsner, 2017). During the explanation phase, unwanted noises might affect teachers' understanding of the content, which in turn could negatively affect their actual performances.

Considering the fact that not all teachers have the same quality of eye-sight, there might be ones who have visual impairments that could hinder their grasp of the content delivered (Fox, 2015). That is, when there are a lot of teachers in a PD session, some of them would have to sit in the back rows as it would be impossible to fit all teachers in the first row, where they can see the content if there is any. Knowing the fact that written and oral exams are scored by addressing to a set of criteria and sample examinations, in a PD session where these are projected, it might be burdensome for some teachers to see or hear the content properly.

Another problem that could arise in face-to-face PD programs can be the reluctance of asking questions to the trainer as the delivery of content takes place in real time. Some instructors, for instance, might feel unwilling to ask questions when they miss or do not understand a particular part in a crowd of teachers. This can happen because there might be more experienced teachers in the session, and asking questions might make teachers feel bad (Salehi, Strawderman, Huang, Ahmed, & Babski-Reeves, 2010).

Lastly, in a face-to-face standardization session, the number of samples to be analyzed and discussed might be much less than a web-based one due to the time limitations of in the workplace (Bennett-Levy, Hawkins, Perry, Cromarty, & Mills, 2012). That is, it might not be possible to keep all instructors at a session for hours since teachers' schedules are not the same.

Many institutions have turned to the utilization of online PD programs for the delivery of training content as these platforms are easier, more time saving and cheaper to conduct than a face-to-face PD session (Keller, 2005).

However, the study (Masters et al., 2010) stated that while the utilization of PD programs provide convenience and effective use of time, these programs' success in enhancing teachers' knowledge and performance in real situations is still a matter of dispute.

1.3. Significance of the Study

This study is noteworthy because of its being the first study to evaluate the effectiveness of an online standardization platform solely created to train teachers before they administer oral examinations as raters. Being totally a web-based platform, this tool would diminish the need for face-to-face standardization meetings by providing institutions and trainers more time to focus on other aspects of the curriculum.

The researcher expected that if the platform proved to be successful, it could yield insights for the standardization of other examinations. The platform's design that enables instructors to get training without time and place restrictions might provide a lot of opportunities for teachers since they could reach the platform no matter where they are and when they want as long as they have Internet connection. For instance, teachers would not have to make time for face-to-face sessions as they can complete the training sessions at home, at a cafe or in a park, thus giving them time to focus on their classrooms and other activities.

Moreover, this study might suggest a new way to deal with problems in a work place. Both schools with a high number of teachers and the ones with a few teachers but a lot of students can make use of this platform to effectively meet their institutional needs.

Furthermore, the results of the study might demonstrate how important PD programs are, and the positive outcomes of this kind of training sessions on teachers' actual performances.

Lastly, the study is important since it makes use of educational technology pertaining to PD programs and standardization of productive examinations thus enabling teachers to become better suited and ready for the advancements in their fields.

1.4. Purpose of the Study

The primary purpose of the study is to evaluate the effectiveness of an online standardization platform solely created for the training of teachers that will administer oral exams as raters. Also, the study will try to find answers to the perceptions of teachers trained on this platform and the ones trained in a face-to-face standardization training with the help of a semi-structured interview in order to gain insights about both standardization mediums.

1.5. Research Questions

The main objective of this study was to explore the effectiveness an online standardization platform specifically designed to train teachers that will administer oral exams as raters. Regarding this aim, this study attempted to answer the following research questions.

RQ1: How effective is an online standardization platform specifically designed to train teachers that will administer oral exams as raters?

- **a.** Do raters trained on an online standardization platform score oral exams consistently within their group?
- **b.** Do raters trained in a face-to-face standardization training score oral exams consistently within their group?
- **c.** Is there a significant difference between the scores of oral exams given by the raters trained on an online standardization platform and the ones trained in a face-to-face standardization training?
- **d.** Do raters trained on an online standardization platform and the ones trained in a face-to-face standardization training apply oral exam criteria consistently in an actual oral examination?

RQ2: What are the views of the participants about the standardization trainings they received before an official oral examination?

- **a.** What are the views of the participants about an online standardization platform specifically designed to train instructors that will administer oral exams as raters?
- **b.** What are the views of the participants about a face-to-face standardization used to train instructors that will administer oral exams as raters?

1.6. Definition of Important Terms

Distance Education is a kind of instruction where learners and instructors are separated physically except for the meetings in person about the course content and projects. It provides opportunities for student interaction online and offline, and student independence in which learners are able to individualize their learning process (Faibisoff & Willis, 2010).

Web-based or Online Learning is defined as a type of formal instructional process where the teaching or training is carried out when the learners and trainers are not present in the same place, and the communication among learners and trainers is utilized by means of Internet (Bhagat, Wu, & Chang, 2016).

Professional Development refers to the development and improvement of knowledge, qualities and skills required to execute tasks in professional work life. By means of professional development, people in a business (Murphy & Calway, 2010) , school (Hughes, Morrison, & Dobos, 2018), hospital (Sinclair, Fitzgerald, Hornby, & Shalhoub, 2015), court (Hou, Horng, & Chen, 2016), and several other professional work areas have the chance to enhance their professional skills, acquire new qualifications, improve their knowledge in their fields, and develop their intra and inter-personal skills (Murphy & Calway, 2010).

Teacher Professional Development is defined as a “long-term” and “inquiry-based” type of professional development where teachers are active learners in their learning processes so that they understand their ideas are given importance. Teachers have the opportunity to experience teaching practices and strategies provided by teacher trainers. Furthermore, teachers have chance to critically reflect upon their experiences and take actions accordingly so as to have a strong and meaningful learning outcome (Jao & McDougall, 2015).

Web-based or Online Teacher Professional Development refers to the instruction where teachers improve their knowledge in their fields physically separated from teacher trainers. Teachers in this type of professional development are able to communicate their ideas and experiences both online and offline. Web-based or online teacher professional development can be realized synchronously in which trainings occur in real-time, asynchronously where trainings are performed at different times, and use of a hybrid or blended method where online trainings are added to conventional face-to-face trainings as a supportive role (Bates, Phalen, & Moran, 2016).

Rater refers to the person who assesses and evaluates exam performances of students by addressing to a set of grading criteria (Wang, 2014).

Standardization or *Rater Training* refers to the process where raters are given training on how to assess performances of students and interpret the scoring rubric (grading criteria) effectively in order to increase intra and inter-rater reliability among multiple raters (Shohamy, Gordon, & Kraemer, 1992; Weigle, 1994)

Interrater Reliability or Agreement is defined as the degree of agreement between or among raters while scoring subjects such as students. Although interrater reliability and interrater agreement are seen as distinct features, this study used them interchangeably by addressing to the literature given (Jonsson & Svingby, 2007; Kottner et al., 2011; McHugh, 2012; Rohner & Katz, 2004).

CHAPTER 2

LITERATURE REVIEW

This chapter provides detailed analysis of themes regarding this study in accordance with the studies in the literature. The literature review includes the themes of web-based learning, professional development, testing and assessment, testing speaking, and building a web-based standardization platform to train raters for oral examinations.

2.1. Web-based Learning

The advent of the Internet has altered how people communicate, work and most importantly how they learn. There have also been several technological developments pertaining to education in the last few decades (Ragan, 2017). Lim (2002) defined web-based learning as a way to enhance learning through network technologies contributing to the communication among learners. Web-based learning, which is a sub-branch of distance education, has become a new trend both for educational institutions and other corporations (Lim, 2002; Myers et al., 2004). Bartley and Golek (2004) suggested that universities today offer web-based learning programs to reach learners from many parts of the world. The advances in web-based learning aim to support learning in several ways such as boosting cooperation among learners (Kirschner, Jochems, & Kreijns, 2005), increasing learners knowledge in various topics (Lim, 2002), strengthening previously acquired knowledge (Edwards, Rule, & Boody, 2017), and observing learners' progress (DeLoose, Unger, Zhang, & Moseley, 2009).

2.1.1. Reasons to Endorse Web-based Learning

There are several aspects affecting the adoption of web-based learning as a teaching model. These aspects are having reasonable costs, enhancing users' learning experiences, offering better instructional design, and providing flexible learning opportunities for users (Bennett-Levy, Hawkins, Perry, Cromarty, & Mills, 2012). The advantages of adopting web-based learning are given below respectively. Training learners on web-based platforms is much more profitable for institutions and corporations since training expenses are reduced to a great extent (Fleischmann, 2018; Flowers, White, Raynor, & Bhattacharya, 2012; Jung & Rha, 2000; Khan, 2002; Powell, 2000). Jung and Rha (2000) and Mayadas, Bourne, and Bacsich (2019) argued that being a cost-effective option, web-based learning enables both academic institutions and other sectors to reduce their expenses regarding learners' travel and accommodation. A wide range of multimedia and learning platforms have proven to be effective in increasing participants' learning experiences, thus helping them get higher grades on their schools (Dewhurst, MacLeod, & Norris, 2000; Enriquez, 2010), be better problem solvers and have better technological skills (Jonassen, Peck, & Wilson, 1999; Larreamendy-Joerns & Leinhardt, 2006). It was stated that technological improvements have helped educational institutions to create web-based learning platforms that have appealing design, content supported by enhanced visuals, tutorials, feedback, and interactive content (Harasim, 2000; Khan, 2002; Powell, 2000).

A number of studies suggest that learners have the chance to obtain knowledge by arranging their own study schedules, accessing training platforms anywhere they wish without time and place restrictions (Coole & Watts, 2013; Fazlollahtabar & Yousefpoor, 2009; Fleischmann, 2018; Harasim, 2000; Khan, 2002; Larreamendy-Joerns & Leinhardt, 2006; McMillen, Hawley, & Proctor, 2016; Scagnoli, 2009; Selwyn, Gorard, & Furlong, 2005; Upton, 2006; Villar & Alegre, 2007).

2.1.2. Concerns about Web-based Learning

In contrast to several benefits of adopting web-based learning platforms, there have also been doubts whether to embrace web-based learning in educational institutions and other corporations. The reasons why educators and employers might not be enthusiastic about web-learning can be summarized as the concerns about users' literacy in technology, social and cultural differences of participants in organizations, and learning styles of participants. All three aspects are explained below respectively.

Although we live in the era of technology and there has been a new technological development almost every day, some people still have problems using technological advents such as computers, mobile phones and the Internet. In spite of its several advantages, Azevedo, Guthrie, and Seibert (2005), Christmann (2017), Debowski, Wood, and Bandura (2001), and Hartley (2002) reported in their studies that participants in a web-based learning environment might be unable to benefit this kind of learning due to lack of technology literacy. Sales and Al-Rahbi (2008) experienced certain problems with participants who had low level of technological knowledge and skills, which hindered the outcome of the study. In addition to the literacy of participants, Giguere and Minotti (2003), Khan (2002), and Powell (2000) suggested that learners should be aided with technological support in this kind of learning environment. It is mentioned that considering technical support factor is of great importance should the organizations aim to offer a web-based training platform to a wide number of participants.

Powell (2000) emphasized that the culture of the institution or the corporation needs to be understood clearly before opting for a web-based training for the people there. Khan (2002) also suggested that some considerations such as socio-cultural differences in the organization, doubts about technology and personal preferences must be taken into account in order to make use of a web-based training platform properly. Giguere and Minotti (2003) addressed some guidelines in their study that the success in web-based training can be achieved by taking learners' preferences and expectations of the training. Gill (2003) reported that adopting a web-based training platform might not prove effective since some participants would like to have an instructor-led training where they are able to communicate with other participants, which may have a greater success rate for the organization. Dedman and Palmer

(2011) expressed that some participants might have negative feelings toward the use of web-based training in their organizations.

Gardner (2011) outlined several learning styles or multiple intelligences which might differ for each individual. While some people might learn better by using their linguistic intelligence, other might do better by using their visual intelligence. Considering this theory, it can be concluded that not every participant on web-based training platform might get the same experience. Lui, Ferrin, Baum, and Randall (2018) promoted multiple intelligences theory by stating that individuals can acquire knowledge much better by using their own way of learning rather than using the ones to which they are not accustomed. Gill (2003) stated that different learning styles might result in different outcomes for the organizations adopting a web-based training platform. That is, a participant with an interpersonal intelligence might not get effective results on a web-based platform.

2.2. Professional Development

The significance of professional development is acknowledged among many different fields such as medicine (Kitzes, Kalishman, Kingsley, Mines, & Lawrence, 2009), engineering (Sunal et al., 2001), arts (Bauer, 2007; Duffy, 2016; Eros, 2013), psychology (Chamorro, 2004; Ciarocco, Dinella, Hatchard, & Valosin, 2016), and education (Griffin et al., 2017; Schaaf, 2018). Hou, Horng, and Chen (2016) expressed that having competency in one’s field, being familiar with the trends (Torff, Sessions, & Byrnes, 2005), enriching communication skills with others, developing cooperation, and paying attention to continuous learning are the key qualities of professionals. So as to cultivate high-quality practices in their fields, professionals such as lawyers, teachers, doctors and scientists need to keep up with changes in their fields and put emphasis on continuous professional development activities (Murphy & Calway, 2010; Webster-Wright, 2009).

Earley and Bubb (2004) defined PD as a process where individuals can develop their competences and knowledge on particular areas both in formal and informal contexts.

Desimone (2009) indicated that people might take PD activities voluntarily or these activities can be obligatory. Moreover, PD practices may be covered through individual work or in co-operation with other staff in the organization. PD can be applied in various models such as action research, reflective practices, and observation (Patton, Parker, & Tannehill, 2015), team teaching (Canaran, 2017), and lesson study (Bayram, Altug, Dereli, Yildiz, & Uzun, 2017). PD is a life-long learning framework where individuals enhance their knowledge and skills through attending job-related activities (Houle, 2006; Penuel, Fishman, Yamaguchi, & Gallagher, 2007; Johnson, 2014), performing collaborative projects (Hill, Beisiegel, & Jacob, 2013), understanding organizational policies, doing reflective practices, and being better problem-solvers.

2.2.1. Teacher Professional Development

After embarking on their careers, teachers could experience troublesome situations such as having overloaded teaching schedules, imprecise expectations from the administration, insufficient support from managers and experienced teachers, seclusion, and undertaking too many responsibilities (Debowski et al., 2001). Especially novice teachers, who are supposed to satisfy the needs of students and administration and improve their teaching skills, face with difficulties regarding the execution of school curriculum and carrying out their profession.

It is emphasized that refining and improving teaching are not easy tasks; therefore, conventional training approaches are sometimes unsuccessful in doing so. As teachers face with a variety of problems in a real classroom environment, the methods, situations, and resources used in training sessions need to be similar to the ones in a real teaching environment. It was also suggested that teachers require continuous training as short-term solutions cannot generate long-lasting effects on teachers' skills. Therefore, in educational settings, administrators and teacher trainers must find ways to practice and manage continuous professional development programs (Hill, 2007).

Teacher professional development has been given great emphasis in the last few decades and several studies (Eun, 2018; Gerard, Varma, Corliss, & Linn, 2011; Hattie, 2008; Houle, 2006; Penuel et al., 2007; Driel & Berry, 2012; Wayne, Yoon, Zhu, Cronen, & Garet, 2008; Webster-Wright, 2009) suggested that TPD activities could demonstrate an increase in students' success in educational settings. It has become an essential part of educational institutions and teachers who strive for improvement in teaching practices and student accomplishment (Hardy, 2016). For educators, PD offers opportunities to acquire and develop content-specific knowledge in their fields where they can pursue both personal achievement and student success. Moreover, Hardy also emphasized that practicing PD activities might yield to better teaching conditions and improve the quality of education in the organization. If performed properly, PD programs can cultivate rewarding outcomes for both administrators and educators in institutions eventually leading to student achievement (Loeb, Miller, & Strunk, 2009). Therefore, in order to be able to benefit from PD programs, school principals and policymakers need to take some guidelines into consideration. Patton et al. (2015) stated that the following guidelines are vital for educational organizations to attain effective results from PD programs.

Haney and Lumpe (1995) and Poekert (2011) argued that the needs, expectations and beliefs of teachers should be taken into consideration if effective outcomes are sought from PD activities. It has been noted that when teachers are aware of insufficient teaching practices or students performing poorly, they might have motives to enroll in PD programs (Saka, 2013). That is, understanding of teachers' needs and expectations is likely to motivate them to participate in PD activities.

Canaran (2017) and Opfer and Pedder (2011) stated that PD programs might yield more effective results if there is collaboration among teachers. Teachers from the same major and working experience in an organization have demonstrated better results in terms of their learning and student achievement. Furthermore, teachers working together on problems create a community where they are able to propose solutions and strategies, which increases the quality of PD for all teachers in the community.

According to Lin and Chiu (2019), PD should create a learning environment that encourages continuous learning for teachers. In this context, teachers become learners that will reflect on, modify and improve the quality of their teaching in a continuous process (Guskey, 2002).

According to Patton et al. (2015), teachers might benefit from PD programs much more positively when there is a direct relation with the content of the program and the application of it in a real classroom environment. It is emphasized that activities such as workshops where teachers can make use of hands-on work can be decisive in student achievement. It is also asserted in their study that activities involving movement and discussions such as presentations and group conversations can be more effective for teachers rather than being passive receivers of information.

In the study conducted by Bezzina (2006), it was discovered that teachers need teaching resources and support of mentors and other experienced teachers. This way, teachers believe that they have a chance to build up on their existing knowledge and use it in their own classrooms. The training related to classroom teaching can bring about positive outcomes when teachers can find enough time to reinforce their knowledge by sharing experiences with other teachers in the organization. Hill (2007) mentioned that the objectives and materials in PD curriculum are likely to be effective in increasing student success since teachers will be exposed to the same or similar resources used in their own classes.

TPD can be realized by means of numerous options such as mentoring (Tondeur, Forkosh-Baruch, Prestridge, Albion, & Edirisinghe, 2016), observation of peers (Zhang & Elder, 2011), workshops (Wayne et al., 2008), self-reflection (Baumgartner, & Hsi, 2019), action research, lesson study (Pella, 2015), team teaching (Canaran, 2017), and videos (Melber, Cox-Petersen, Berg, & Enochs, 2005).

2.2.2. Web-based Teacher Professional Development

Along with the developments in web-based learning and the growing numbers of web-based learning platforms, educational organizations have also started to launch web-based PD programs where teachers could be delivered effective materials to enhance their knowledge in their fields and teaching methods continually (Anderson & Henderson, 2005). It was argued that despite numerous studies about the importance of continuing PD, realizing a systematic continuing PD has turned out to be ineffective due to monetary reasons and interruption of teachers' teaching schedules. As a result, school administrators have started initiating web-based PD as an idea to accomplish practical PD with reasonable prices and flexible time options. Diaz and Bontenbal (2001) also indicated that educational organizations could benefit both web-based and conventional TPD programs by combining them and providing rational time allocations for each since teachers in numerous organizations cannot find time for conventional PD programs.

Web-based TPD can be realized in several different methods such as tele-mentoring, web-based learning forums, virtual worlds, video streams, all of which have been observed presenting solutions for TPD programs. In order to get the best results out of web-based PD, some key factors are needed to be taken into account by teachers and organizations wishing to practice their PD programs on a web-based platform. These aspects are (a) providing teachers with relevant materials to their actual teaching environment, (b) creating an organized PD environment, (c) supporting teachers with on-time feedback on their performance, (d) interactive options, and (e) offering teachers a flexible learning environment in terms of place and time (Levin, Waddoups, Levin, & Buell, 2001).

School principals should also focus on some aspects before opting for web-based teacher professional development. It was mentioned that school administrators need to be aware of the technological literacy levels of the teachers to enroll in TPD activities. It is suggested that teachers would work better in an environment where they are aware of their competences and confident about the tasks provided. Therefore, knowing about the profiles of teachers concerning technological literacy is one of the key factors for principals before making a decision. According to Venkatesh and Davis (2000), another significant aspect is the beliefs and attitudes of teachers towards web-based TPD. It was mentioned in their study that teachers have the opinion of deciding whether taking an online TPD would benefit them or not. That is, enrolling in a web-based TPD is voluntary. Teachers tend to weigh the positive and negative sides of particular actions and decide whether it is a good idea to do it or not (Ertmer, 2005). Moreover, teachers' perceptions have a direct effect on their acceptance of web-based TPD. In other words, whether the adoption of a web-based TPD will improve teachers' knowledge and teaching skills would be a fundamental factor for teachers in endorsing a web-based TPD program. It is emphasized that school administrators need to carry out a needs analysis in order to be able to conclude whether a program is necessary for the staff in the organization. Training is performed to enhance or shape teachers' knowledge in their fields, skills and abilities regarding their personal lives and teaching methods. Hence, teaching staff should be sure that embracing a web-based TPD program would provide them with these achievements.

There have been various web-based TPD programs targeting different types of teachers. Latchem, Odabasi, and Kabakci (2006) created a web-based TPD platform aiming to train computer teachers who have just started their teaching careers. The platform included three sections in which teachers are given training regarding teaching, rights and duties of teachers, and the structure of schools.

Villar and Alegre (2007) examined the effects of two web-based PD courses that provided support for junior faculty at a university. The study aimed to find out whether these two courses had positive impressions on the teaching staff that was exposed to support systems, workshops, and professional guidance by an experienced mentor professor. The results of the study showed that participating teachers had positive attitudes about web-based TPD, and they reported that they had broadened their horizons regarding teaching scientific curriculum.

Hur and Hara (2007) conducted a study where they attempted to learn about the aspects regarding a web-based community for K-12 teachers. The study was carried out on a platform called INDISCHOOL in Korea. Interviews with participants, observations and entries on the platform were thoroughly analyzed by the researchers. The findings of the study demonstrated that participating teachers believed the platform enhanced their teaching skills and student success. The teachers also had positive attitudes towards the platform where they were able to have autonomy over their learning and understand the benefits of collaboration with other teachers.

Masters et al. (2010) carried out a project which was a part of “e-Learning for Educators Initiative”, which aimed to offer high-quality web-based professional development in the field of English Language Arts in eight different states in the United States. The study consisted of three workshops that lasted over seven weeks and required about five hours of attendance of teachers on the platform. Each workshop included reading assignments, online practices and discussions. The findings of the study suggested that teachers attained valuable knowledge and experiences about their teaching practices.

Arikan (2006) studied achievement, recalling and opinions of pre-service computer teachers on a web-based TPD platform. The results of the study demonstrated that the participants on the web-based TPD platform had more positive attitudes towards than the ones who were trained face-to-face. The findings of the study also showed that participants favored the web-based TPD platform as they had the flexibility in terms of time and place. However, the participants mentioned some constraints regarding the web-based platform such as inadequate interaction and insufficient feedback as opposed to conventional PD.

2.2.3. Web-based TPD versus Face-to-face TPD

Web-based TPD programs have become much more popular in the last few years as face-to-face TPD programs pose some constraints for trainers such as the problems in scheduling and lack of variety in offerings (Kleiman, 2004). As mentioned earlier, web-based training platforms have several advantages over its face-to-face counterpart. James (2002) mentioned these aspects of web-based PD platforms as being efficient, cost-effective, and flexible. According to The Web-Based Education

Commission (2000), web-based TPD has many benefits because of the fact that teachers have the opportunity to exploit high-quality learning resources and interact with experts on their fields that they normally would not be able to communicate with in face-to-face PD programs.

Scott, Feldman, and Underwood (2016) argued that web-based TPD programs might have more or the same effect as face-to-face TPD programs. However, they stated that some serious topics such as depression and trauma which would require more human interaction. Therefore, there would be a demand for more face-to-face discussion. This way, they expressed that using a web-based TPD in this context may not yield as efficient results as a face-to-face TPD.

Scott et al. (2016) also added that teachers with a high level of technological literacy would probably choose web-based TPD programs as they are not afraid of using technological advents. This way, their confidence would inevitably generate greater success rates than teachers who have problems with using technology. As an example to this notion, the findings study conducted by Kao, Tsai, and Shih (2014) demonstrated that teachers who are confident in using technological devices and the Internet are more likely to adopt a web-based TPD program.

While technology may have positive effects on TPD programs’ success, some concerns regarding web-based TPD programs can impede the use of web-based TPD effectively in educational organizations. Huai, Braden, White, and Elliott (2009) asserted that if teachers are not willing or ready to adopt a web-based TPD program, the success rate could decrease considerably. They also furthered their opinions by stating that absence of a real trainer in the platform might bring about some problems as well. For instance, whereas hyperlinks on the TPD program can help teachers to find information easily and in accordance with their own pace, they can also confuse teachers as they might not be guided throughout the program. Therefore, the web-based TPD program should be created in a way that teachers will not have problems related to guidance.

Web-based TPD programs can yield to effective outcomes when teachers, who are going to participate in them, are given training on how to use the platform, and what to do on the platform prior to the commencement of the training (Bohnenkamp & McMahon, 2001). Additionally, according to Putnam and Borko (2000), the multimedia when supported by real-life examples such as student presentations, mentor videos, and sample student examinations can generate greater achievement rates for teachers on web-based TPD programs.

Glomb, Midenhall, Mason, and Salzberg (2017) concluded that teachers favored web-based TPD training instead of face-to-face TPD since they were able to cultivate more knowledge and experience from online mentoring facilities and online learning communities where they could share their ideas with other teachers and have discussions about the content of the training. In addition to this, Erickson, Noonan, and Mccall (2017) mentioned that web-based TPD could boost the interaction and cooperative work among teachers. They also furthered their notions about web-based TPD by stating that web-based TPD programs could support novice teachers in retention of content-knowledge.

Masters et al. (2010) argued that while there are many advantages of embracing web-based TPD programs such as flexible time options, cost-effectiveness and ease of reaching larger audiences, there are still questions whether web-based TPD could significantly alter and develop teachers' teaching skills and boost student achievement as opposed to traditional TPD programs that are conducted face-to-face.

Baran and Cagiltay (2006) carried out a study about a web-based TPD program. According to the results of the study, the teachers expressed their opinions by emphasizing that the web-based TPD program had several advantages such as flexibility regarding time and place, and setting of the program. However, they also had some difficulties with the web-based TPD program. For instance, teachers stated that the training lacked coherence with their teaching context and they had technical problems during the training but received no support. Teachers also added that web-based TPD programs could be very beneficial for them if these problems are fixed and the programs are developed in accordance with their expectations from the training.

On the other hand, there have also been some studies in which lack of face-to-face interaction proved to be unsuccessful. For instance, Howard and McGrath (1995) reported that pre-service English Language teachers participating in an online PD had much lower results compared to their fellows trained in a face-to-face PD program.

Powell, Diamond, Burchinal, and Koehler (2010) carried out a study where they examined the similarities and differences between web-based TPD and face-to-face TPD. Their objective was to find out whether any of the modalities had a significant effect on teachers' teaching skills and student achievement. The results of the study suggested that teachers demonstrated improvement in their teaching skills and student achievement in both of the modalities. However, the results did not favor any of the modality as being superior to the other. The researchers concluded their study by emphasizing that web-based TPD might not be better than face-to-face TPD; however, it can be a potential alternative to face-to-face TPD.

In another study by Fisher, Schumaker, Culbertson, and Deshler (2010), some teachers were randomly selected for a training course where they were required to learn about a concept mapping method to enhance students' learning. Then teachers were grouped randomly into web-based TPD or a face-to-face TPD. The web-based TPD included a CD-ROM, which consisted of resources and lesson plans of the course, and discussions with mentors and other teachers. The face-to-face TPD had the same content, but the resources and lesson plans were distributed in-person, and the discussions were conducted in a real classroom environment. The results of the study showed that there were no significant differences in the knowledge of teachers participating in a web-based TPD and face-to-face TPD. However, it was emphasized that web-based TPD could be a promising alternative to face-to-face TPD.

Holmes, Polhemus, and Jennings (2005) studied a blended model TPD program called CATIE which stands for "*Capital Area of Technology and Inquiry in Education*". In their study, they created a content and question based TPD program to support K-12 teachers. By taking continuity and fiscal issues into consideration, the researchers settled on giving trainings with a blended model. CATIE contributed to the development of K-12 teachers in several ways. Firstly, there were mentors working online replying teachers' questions, which eliminated the need for on-site mentors and all participants had a chance to get information from different mentors. Also, teachers could communicate with other teachers from different institutions and have the chance to share ideas with each other. This way, they were able to gain new insights about their fields and the use of technology in their professional lives. Lastly, being exposed to different modes of materials and resources, teachers participating in CATIE demonstrated increased levels of content knowledge, collaboration and communication.

2.3. Testing and Assessment

Testing and assessment are terms that are often thought as synonyms of each other. While assessment means the evaluation of the students' achievements and processes of their learning, testing can be defined as the procedure conducted in a periodical way in curriculum of the institutions. In addition, testing is applied at certain times and used by means of administrative objectives whereas assessment is a continuous process (Parkes, Zimmaro, Parkes, & Zimmaro, 2018). The study by Leung and Lewkowicz (2010) also argued that assessment can be accepted as a superordinate type of all types of assessment found in the literature. However, testing is defined as a particular type of assessment. That is, although assessment can be utilized in several ways regardless of being formal, informal, deliberate, or unintentional, testing requires curricular or standardized means of implementation. There are two kinds of assessment in the literature, which are formative and summative assessment types.

2.3.1. Summative Assessments

Drouin (2010) defined summative assessments as tools implemented in a formal context to test students in order to be able to report their achievements. This way, school administrators and teachers are likely to know whether the course objectives and expectations from the students have been met or not by looking at the scores of the students. Moreover, educators can adopt or develop their current ways of teaching and testing so as to enhance what their students are capable of doing and develop a more effective instructional setting (Sharkey & Murnane, 2006). Standardized tests play a major role in assessment since they are one of the most commonly used assessment types in educational settings. Standardized tests such as TOEFL and IELTS have been dominating the field of English language for many years (Leung & Lewkowicz, 2010). It was also argued in the study that there have been attempts to develop teacher-based assessment types as debates such as providing unauthentic information regarding the standardized tests aroused in the field of English instruction. Due to these debates, these standardized tests have still been used as a way to demonstrate particular achievements of the students. Qu and Zhang (2013) presented some benefits and concerns about summative assessment types as shown below. The first benefit of summative tests was shown as a way to develop or change curriculum lessons by taking the numerical data extracted from summative tests into consideration. Therefore, summative tests could play a significant role in enhancing the quality of teaching in educational settings. The second benefit was that students would be able to see the areas or lessons where they have difficulties. Thus, they might attempt to solve their problems by contemplating on their exam scores. Also, they might be able to get help from their instructors in order to improve their studying skills by changing their habits of studying. As seen in Table 2.1., some major summative assessment types can be midterm and end-of-term exams, unit tests, standardized tests such as SAT, TOEFL, IELTS, etc. (Hoover & Abrams, 2013).

2.3.2. Formative Assessment

Volante and Beckett (2011) defined formative assessment is continuous and occurs in a lesson or study such as reflection journals and self-assessment surveys. Lorraine M. Baron (2016) described formative assessment as a tool to help teachers to shed light on our teaching practices through evaluating and helping students to improve their skills. According to the study by Brookhart (2013), formative assessment can help learners to have self-regulation in the course of their learning processes. It was argued that students would be able to guide themselves into the correct direction to achieve their educational goals if they are given a chance to monitor their own learning process, make use of the feedbacks they get from their teachers, and have solid learning goals. Formative assessment benefits from teacher feedbacks to modify teaching procedures in order to meet the needs of students throughout a teaching period. So, as to enhance teaching at a school, formative assessment can be utilized in order to grasp the connection between students and teaching (Steward, Mickelson, & Brumm, 2004). Some benefits of formative assessment are described as boosting low-achiever student's confidence and their results in summative tests, improving student learning, bonding students and teachers in terms of educational feedback and exchange of ideas, enhancing the quality of the instruction in educational institutions (Steward et al., 2004).

Main types of formative assessment are questions asked by teachers, feedbacks, self-assessment, peer reviews (Volante & Beckett, 2011), journals, surveys (Shirley & Irving, 2015), term papers, portfolios (Nolen, 2011), diagnostic tests, exams that are not graded, self-check documents (See Table 2.1.) (Russell & Blake, 1988).

Table 2.1. *Summary of Assessment Types Used in Summative and Formative Assessment*

Summative Assessment	Formative Assessment
Midterm exams	Questions asked during the lessons
Final exams	Verbal and written feedbacks
Standardized tests	Peer reviews
Unit tests	Journals
Diagnostic tests	Surveys
	Term papers
	Portfolios
	Diagnostic Tests
	Not-graded exams
	Self-check documents

2.3.3. Second Language Testing

Language tests possess a very significant role in education. They might have various effects that could go beyond what they aimed to achieve. They have an important impact on learner motivation and dedication. Shohamy (2006) stated that a positive testing experience could lead to an increased level of motivation and success while a negative one could hinder the development and progress of students. Therefore, language testing is of great significance as it is related to both the curriculum of the schools and the achievement of students. Bachman (2000) argued that language tests are implemented so as to learn about the proficiency levels of students in particular skills such as grammar, vocabulary, listening, reading, writing, and speaking. It was also expressed in the study that when tests are implemented conventionally, they could help assess performances of learners. The scores extracted from these tests could help institutions to understand their students' particular language skills and improve the teaching quality at the school.

2.4. Testing Speaking

Amongst the various test types attempting to assess listening, reading, grammar, writing and vocabulary skills of learners, testing speaking is relatively new in the field of language testing (Fulcher, 2014). With the movement from traditional methods of instructions into a more communicative approach, assessing speaking started to receive a great deal of emphasis. Thus, educators felt the need to improve speaking skills of learners in order for them to survive in a communicative learning environment (Larson & Larson, 2019; D. Lin & Liu, 2018). Luoma (2004) stated that testing speaking performances of learners was a quite difficult job since there might be several reasons affecting the overall oral performance of the learners. Furthermore, learners find speaking as the most challenging skill while learning a foreign language.

2.4.1. Methods for Testing Speaking

In the context of English as a second language, speaking skills of learners play an important role in order to communicate with peers and teachers in the classroom.

Therefore, learners must have the ability to comprehend what has been said and make logical utterances accordingly. For these reasons, communication skills bear great significance for learners to be successful in ESL contexts (Murphy, 2006). Types of assessment for speaking skills can be decided according to the objectives of the school. For instance, in a Business English classroom, learners can be tested by oral presentations (Murphy, 2006). Another method to test speaking performances would be unintegrated speaking tests where test takers are not given any input before they start speaking (Brown, Iwashita, & McNamara, 2005). As opposed to unintegrated speaking tests, language testers could also utilize integrated speaking tests where test takers are provided with oral or written input before they commence to speak (Huang, Hung, & Plakans, 2018). Speaking performances of learners can also be tested by full automated software or online such as Pearson's Versant tests (Bernstein, van Moere, & Cheng, 2010). Speaking test might also cover tasks such as topics and situations that test taker are supposed to talk about during the exam (Fulcher, 2014). These tests could also apply reading aloud, presentations, and role play (Zhao, 2013). Lastly, test takers can be tested by means of verbal questions by the raters and describing visuals (O'sullivan, Weir, & Saville, 2002).

When rating students' spoken performances, raters can make use of a variety of criteria based on different components of the language (Babaii, Taghaddomi, & Pashmforoosh, 2016), which are listed below.

- Accuracy in using the target language grammar
- Using a variety of appropriate lexical items
- Cohesion and clear use of ideas
- Fluency in the target language
- Being relevant to the topic
- Discourse markers
- Pronunciation
- Timing

Grammar and Vocabulary

Römer (2017) discussed that grammatical knowledge and lexical proficiency were assessed separately from each other in the past. Therefore, it many of the past criteria used to evaluate oral performances, these were assessed separately. With regard to these reasons, several rating scales did not test these components together. However, as there have been new approaches in the field of language testing, researchers started to treat these two components in one single category. It was argued that verbal productions such as sentences and phrases were actually a combination of lexicon and grammar in the language (Ellis, Römer, & O'Donnell, 2016). Thus, these two components were started to be used in various fields of language testing.

Discourse Markers

The use of discourse markers enables learners to take breaks during their speech. As it is seen in the spoken performances of native speakers, hesitation is a natural action used by people for several reasons. Utilization of discourse markers such as uh, erm and um provide test takers with enough time to plan their next utterances by organizing their ideas on their minds. Apart from these, discourse markers such as well, so, let me think, etc. are seen as a way to be more cohesive during the spoken performances (Römer, 2017). Therefore, it can be said that these components could be the sign of speaking proficiency as they are often used by native speakers of the language.

Pronunciation

Isaacs, Trofimovich, and Foote (2018) argued that pronunciation is a vital component of speaking rubrics. By evaluating the ability of speakers regarding pronunciation, it could be demonstrated how developed the speaker's comprehension and production in the target language. However, it was argued by Isaacs & Harding (2017) that the objective of the raters should be towards assessing the speakers' being intelligible in the target language rather than the accent regarding the pronunciation of test takers.

2.4.2. Importance of Rater Training in Testing Speaking

Standardization or training of instructors or raters who are going to score productive tests such as writing and speaking can be defined as the process where instructors are provided with sample examinations and grading criteria in order to agree on a shared interpretation of how exams are scored in accordance with the criteria (Wigglesworth, 1993).

Elder, Barkhuizen, Knoch, and von Randow (2007) and Wind (2019) asserted that standardizing instructors for tests such as speaking exams is fundamental for the instructors themselves and schools because the instructors who are trained for these exams are able to refresh and reorient themselves regarding the exam criteria and their own progress. While doing this, they also have the chance to exploit training materials and sample examinations, which allow them to be ready for the tests. With the implementation of rater training, raters could be more consistent in their scoring of examinations. However, it is noteworthy that the effects of training might not last for a long time after a training session; therefore, standardizing raters continually before examinations play a vital role in reestablishing and internalizing the grading criteria and consistent scoring (Fahim & Bijani, 2011; Wang, 2014). It was also stated in Wang's study that instructors who are going to be appointed as raters should be given trainings regularly so as to obtain experience in the examinations, which is highly likely to make them confident and consistent throughout their rating experience. Knoch, Fairbairn, and Huisman (2016) argued that standardizing raters in training sessions is vital before assigning them as raters in tests due to the fact that raters are likely to differ in their subjective judgments and interpretations of the grading scales. These differences might arise from rater factors such as rater leniency or severity, inadequate grasp of the criteria, and rater bias (Bijani, 2018; Kondo, 2010).

Kondo (2010) argued that standardizing raters is often implemented so that there could be a certain level of agreement among raters regarding their scores. Kondo also added that achieving an agreement level where all raters appoint the same scores is unfeasible with regard to severity of raters. However, the standardization trainings enable raters to become more consistent with themselves and other raters. The factors affecting the rater agreement might not be totally eradicated, but huge score differences among raters could be avoided thanks to the utilization of standardization sessions. Tajeddin and Alemi (2014) discussed that assessing productive exams such as speaking could bring about serious problems stemming from rater differences. Therefore, conducting standardization programs could help raters to increase their self and inter-consistency (agreement) levels in assessing performance exams, which would eliminate problems resulting from rater variability such as severity, leniency, and interpretation of rating scales. Tajeddin and Alemi also stated that rater training could yield to fruitful outcomes as score differences among raters are decreased, confidence level of raters and the inter-rater agreement among raters are increased after training.

Contrary to positive findings, Congdon and McQueen (2000) discussed that rater differences such as severity in scoring might exist even though raters have been trained. Lunz and Stahl (1990) reported differences of agreement among raters a short time after the training session stating that the standardization was not effective. Similarly, Eckes (2008), Lumley (2005) and Weir (2005) mentioned that there could still be differences in agreement levels of raters after extensive training sessions, which could arise from different interpretation of the grading scale and rater bias. Hamp-Lyons (2012) asserted that standardization sessions might not be persistent in decreasing the inconsistency levels among raters due to the differences in raters' background such as being experiences or novice.

Despite some drawbacks, rater training plays a significant role in increasing the agreement levels among raters and intra-rater consistency. Terry and Hughes (2006) expressed that raters are able to score more consistently thanks to the standardization procedures as they are provided with extensive information about the grading criteria and how to use it while assessing student exams. Furthermore, consistency or agreement levels and fairness among raters to a great extent by the utilization of standardization sessions (Bachman & Palmer, 1996; Brown, 2007; Myford & Wolfe, 2000). Also, İlhan and Cetin (2014) argued that rater training might yield to persistent levels of inter-rater agreement if multiple raters are included in scoring sample exams and while assessing real student exams. Tajeddin, Alemi, and Pashmforoosh (2011) found in their study that raters show increased levels of agreement among themselves following standardization sessions regarding the application of grading criteria and intra-consistency. In conclusion, several studies (Elder et. al, 2005; Iwashita, Brown, McNamara, & O'Hagan, 2008; Knoch et al., 2016; Lumley & O'Sullivan, 2005) suggested that implementation of rater standardization is of great importance so as to increase the agreement among raters while assessing oral examinations.

2.4.3. Standardizing Raters Online vs Face-to-face

Knoch et al. (2016) reported that adopting an online training for standardizing raters might be more beneficial for institutions than its face-to-face counterpart as a high number of raters to be trained might hinder the utilization of face-to-face training regarding the meeting and logistics, conducting a face-to-face standardization with a lot of raters might not be effective, and raters differ from each other in terms of the length of time while analyzing sample exams, evaluating, and scoring them.

Hamilton, Reddel, and Spratt (2001) carried out a study in Hong Kong Polytechnic University for supporting the English Language center there. The study was conducted with instructors who received online or face-to-face training about the oral examinations implemented in the school. The findings of the study suggested that the instructors trained online demonstrated positive feelings about the online training in terms of practicality and ease of access (Upton, 2006), but the utilization of an online training platform requires a section where raters are able to discuss their scores with the other raters and receive feedbacks (Gill, 2003; Tynjälä & Häkkinen, 2005). Similarly, Bijani (2018) mentioned that receiving feedback in online standardization is vital for raters as it could diminish rater bias and increase the level of agreement while scoring student exams.

Erlam, von Randow, and Read (2013) offered a web-based training in order to standardize raters as part of the "Diagnostic English Language Needs Assessment (DELNA)" program at the University of Auckland. Their idea was to increase the intra and inter-rater agreement among raters. The results of the study demonstrated that online training bears potential to standardize raters for future examinations. The raters who participated in the study also showed positive feelings towards the use of online training due to practicality and ease of use. Similarly, employing an online training could be more effective as it could diminish some problems of face-to-face training such as being time-consuming and difficulty of gathering a number of raters in the same place (Jung & Rha, 2000; Powell, 2000)

Glomb, Midenhall, Mason, and Salzberg (2017) concluded that teachers online training of raters could cultivate more knowledge and experience as they could share their ideas with others and have discussions about the content of the training. In addition to this, Erickson, Noonan, and Mccall (2017) mentioned online standardization could boost the interaction and cooperative work among raters, and help novice teachers increase the retention of content-knowledge.

Davis (2016) investigated the effects of online training conducted with TOEFL IBT raters who scored 100 student oral exams before and after the training. Following the analysis of the study, it was found that there had been an increased level of inter-rater agreement and agreement with the pre-established reference grades. It was also suggested that training raters on web is highly suitable because the raters are able to make use of a lot of sample oral exams as opposed to a conventional face-to-face standardization program.

Christmann (2017) reported that raters receiving online training might have different agreement levels from the other raters due to not having enough technological literacy. Therefore, implementing a pilot study would be a good idea prior to the training sessions. Gill (2003) emphasized that some raters might want to take part in a trainer-led face-to-face standardization so that they could exchange ideas about their decisions with other raters. Elder, Barkhuizen, Knoch, and von Randow (2007) reported in their study that raters on the online platform demonstrated low levels of inter-rater agreement and concluded that the raters should have been given a face-to-face training before they embarked on the online platform.

Scott, Feldman, and Underwood, (2016) reported that both online and face-to-face training modalities had the similar results in terms of rater training. Powell et al. (2010) noted that instructors trained in either face-to-face training or online training were not better than the other in terms of the improvement of their rating skills. Both modalities of training had similar effects on the instructors' agreement among each other.

2.4.4. Rater Reliability in Speaking Tests

Rater reliability or rater agreement is a crucial part of a speaking test as the decisions made by the raters during a language test might cause severe problems in terms of validity and reliability of the test being conducted (Davis, 2016). The problem with the rater agreement is that raters are provided with a scoring rubric during the exam even if there are some occasions where there are complex examples of student performances (In'nami & Koizumi, 2016). These challenges might affect the overall score given by the raters if they have not fully understood the criteria and given a different score rather than required (Lumley & Mcnamara, 1995).

However, several studies have shown that raters are able to give consistent scores for the oral performances of learners if they have received an effective training. There have been an increase in the rating performances of raters after receiving the training in many studies (Davis, 2016; May, 2009).

Kang, Rubin, and Kermad (2019) conducted a study in order to see whether training raters before a speaking exam would decrease rater differences in the implementation of a speaking test. Therefore, they worked with 40 raters without training in the study in order to rate sample speaking exams of TOEFL IBT test takers. In the first phase of the study, differences among raters' scores were identified due to some reasons such as background and experience. The second phase of the study involved an online training. The 40 raters received an online training and then they re-scored the exams they had rated before in a random order. The results of the study supported that receiving training before a speaking exam would help raters to grasp the criteria thoroughly, eliminate biased views about test takers, and enhance their rater reliability.

Yan (2014) carried out a study in which rater performances in an English proficiency test were evaluated. The raters who participated in the study received an online training regarding the oral examination of the school. Consistency of raters in scoring, agreement among raters and their use of the scoring rubric were analyzed. The results of the study showed that raters had a positive level of inter-rater reliability (agreement) and the interpretation of the scoring rubric.

As a result, it can be argued that inter-rater reliability or agreement among raters can be increased through training platforms where they can be exposed to sample examinations and gain experience in their fields as raters.

2.5. Building an Online Standardization Platform

2.5.1. Cognitive Load Theory

Paas, Renkl, and Sweller (2004) defined CLT as a theory which deals with learners' interactions with various informational elements and their cognitive processes on their minds. According to their study, CLT is also related to the working memory load and the types of cognitive loads on this particular memory type such as intrinsic cognitive load, extraneous cognitive load and germane cognitive load.

Instructional design principles and CLT are in accordance with each other. By taking CLT into consideration, learning outcomes could be constructed as CLT helps us understand the cognitive architecture of human mind. In this way, training programs with better instructional design and less cognitive load can be utilized.

Takir and Aksu (2012) studied the effect of a training designed by taking CLT into account in a middle-school. The study attempted to find out whether this particular training was more effective than the traditional teaching in terms of students' success rates in algebra. The results of the study showed that the training prepared by focusing on CLT proved to be much more effective than the traditional one.

2.5.2. Instructional Design Methods

Learning through multimedia has a direct relationship with Cognitive Load Theory as CLT is concerned with how different modes of presenting information are related to grasping and using information. Mayer (2012) discussed that modes of presentation such as texts and visuals along with the sensory modalities such as visual and auditory are the fundamentals of instructional design due to the fact that acquiring information fully can be realized by displaying the learning resource by using both text and visuals. That is, presenting information in both text and visuals help learners apprehend the content much better.

In his book, Mayer suggested twelve fundamental methods that affect learning through multimedia. These are (1) *coherence*, which examines whether there is any extraneous (irrelevant) information in the instruction, (2) *signaling,* which is about highlighting important information, (3) *redundancy,* which inspects whether the information interferes with dual-channel theory multimedia learning that supports the idea that information should not be given just for the sake of giving it. For instance, some information could be given better visually without needing another mode of presentation or vice versa. Therefore, simply presenting information in different ways but in the same mode of presentation would eventually cause cognitive overload in learners' minds. An example for this would be putting Turkish subtitles to a video spoken in Turkish. (4) *Spatial contiguity,* which investigates whether the visuals and texts are located close to each other or not. Moreno and Mayer (1999) stated that visual resources should be located close to written resources in order for learners to comprehend the content easily. (5) *Temporal contiguity,* which examines if visuals and related sounds are presented at the same time or not. Showing a picture of micro-processor while the object is being uttered in a video or animation could be an example of temporal contiguity. (6) *Segmenting,* which is about presenting the information in parts or continuously, (7) *pre-training,* which is concerned with whether learners are given training about the concepts in the instruction, (8) *modality,* which investigates whether the visual information is supported by auditory material or vice versa, (9) *multimedia,* which is related to designing the instruction by combining texts and graphics, (10) *personalization,* which is about whether the texts (written or spoken) are given in a formal or informal way, (11) *voice,* which examines whether the instruction is given by a real person's voice or a machine's voice, (12) *image,* which is related to the presence of learners' image on the learning platform.

CHAPTER 3

METHODOLOGY

This chapter includes detailed information regarding the assumptions, research questions, participants and settings, role of the researcher, research design, implementation and procedures, data analysis, trustworthiness and the limitations and delimitations of the study respectively.

3.1. Assumptions

The researcher assumed that all participants selected for the study were computer-literate as they used computers provided by the school in their classrooms all the time to project the course books on the board. Another reason for the first assumption was that all instructors who took part in the study were between the ages of 22 and 30, so the researcher had assumed that they were familiar with technological devices and the internet. It was also assumed that all participants responded to the data collection tools employed by the researcher intentionally since they knew the researcher in person and they were colleagues in the same institution.

3.2. Research Questions

The main objective of this study was to explore the effectiveness an online standardization platform specifically designed to train teachers that will administer oral exams as raters. Regarding this aim, this study attempted to answer the following research questions.

RQ1: How effective is an online standardization platform specifically designed to train teachers that will administer oral exams as raters?

a. Do raters trained on an online standardization platform score oral exams consistently within their group?

b. Do raters trained in a face-to-face standardization training score oral exams consistently within their group?

c. Is there a significant difference between the scores of oral exams given by the raters trained on an online standardization platform and the ones trained in a face-to-face standardization training?

d. Do raters trained on an online standardization platform and the ones trained in a face-to-face standardization training apply oral exam criteria consistently in an actual oral examination?

RQ2: What are the views of the participants about the standardization trainings they received before an official oral examination?

a. What are the views of the participants about an online standardization platform specifically designed to train instructors that will administer oral exams as raters?

b. What are the views of the participants about a face-to-face standardization used to train instructors that will administer oral exams as raters?

3.3. Participants and Settings

Considering the fact that studying an entire population belonging to a particular research area is barely possible, various sampling methods are utilized in order to be able to represent the total number of individuals associated with a specific study area (Creswell, 2013). Among the methods to select the samples, which are probability sampling and non-probability sampling, the researcher opted for using the latter since the setting of the study and the number of the people who could be selected for the study did not allow probability sampling. Furthermore, because of the nature of the study, the researcher decided that the needs and requirements of the study would be met better by means of non-probability sampling. That is to say, the researcher was able to appoint samples pertaining to particular characteristics such as work experience and experience in oral examinations of the institute they had been working.

In the light of these reasons, the researcher had the idea that non-probability sampling was the better option as opposed to probability sampling.

The study employed purposive sampling method as the focus of the researcher was to create two groups of instructors who would be trained on an online standardization platform and in a face-to-face standardization session. Both groups were designated to have both experienced and novice instructors with regard to the institution's official oral examination. Also, as for the number of instructors, both groups were planned to have the same number of experienced and novice instructors. Etikan, Musa and Alkassim (2014) stated that purposive sampling method, a type of non-probability sampling, pose some drawbacks to researchers since it is subjective regarding the selection of the samples. Therefore, it might not necessarily represent the whole population of the researcher's area of study. However, when the researcher's limitations such as being unable to find a larger size of samples, having limited time, and not having enough number of samples with different work experiences and experience in the department's official oral examination are taken into consideration, the researcher decided to employ purposive sampling method as it helped eliminate such limitations while conducting the study. Among the types of purposive sampling, the researcher's decision was to use homogenous sampling since the design of the study was to include samples with similar backgrounds (Sharma, 2017) such as experience in the official oral examination in the school they had been working.

Table 3.1. *Participants of the Study*

	Online Standardization		Face-to-face Standardization	
Gender	Novice*	Experienced*	Novice*	Experienced*
Female	5	4	5	4
Male	1	2	1	2
Total	6	6	6	6

* The terms "novice" and "experienced" refer to the experience as a rater in the official oral examination of the school

The samples who took part in this study were 24 English Language instructors working in the department of foreign languages at a foundation university in Ankara, Turkey. As mentioned before, they differed in terms of their experience in the official oral examination of the institute they had been employed. All experienced instructors had been working in the department for at least four years and had conducted several oral examinations as raters in the school's official oral examination by the time the study was carried out in 8 January 2019. On the contrary, all novice instructors in the study had been working in the department for just five months and they had not had any experience in the school's oral examination either. The research was carried out in a period of 3 weeks from 10 December 2018 to 31 December 2018.

From a total of 24 samples, the researcher assigned 12 instructors (6 experienced and 6 novice) for each standardization type, which were online and face-to-face standardization. While selecting the samples for the study was done by using purposive sampling method, assigning experienced and novice instructors into online and face-to-face standardization was done randomly (See Table 3.1.).

As for the setting of the study, all English language instructors at the department are employed on a contractual basis that is renewed yearly. The maximum workload of instructors is 20 teaching hours per week. In addition to teaching, the instructors are also assigned to tasks such as exam invigilation, marking writing exams and translating official documents. Regardless of their teaching schedule and whether they have classes or not, all instructors work full-time during the week. When they do not teach classes, they also have office hours where they help students with their lessons or carry out extracurricular activities such as speaking clubs, grammar lessons and cultural activities. All instructors are given the chance to pursue academic degrees such as masters and doctorate in other universities; thus, they are given two half-day offs if they are registered to one. Apart from that, all instructors are given a half-day off every week regardless of whether they pursue an academic degree or not.

3.4. Role of the Researcher

While discussing the characteristics of quantitative and qualitative studies, there are some factors to take into consideration regarding the role of the researcher. Yilmaz (2013) argued that in a quantitative study, the researcher and the samples of the study are seen as separate and independent from each other. That is, the researcher bears an objective role in the implementation of the study. Therefore, the researcher in a quantitative study detaches himself from the subjective notions and has an etic role which is the role of an outside and objective view. This results in a drawback where participants are not allowed to share their own feelings towards the instruments of the study and their experiences throughout the process, which means that the subjective interpretations of the samples are ignored (Patton, 2002).

Conversely, qualitative studies are more concerned with the implementation process, interpretations and the subjective interpretations of the instruments of the study. That is, experiences and opinions of the participants are given importance by employing observations and interviews to gain deeper insights of them. Therefore, the researcher's role in a qualitative study is emic, an insider point of view, which enables subjective participation, being partial and use of empathy with the samples (Bergman, 2008; Creswell, 2007).

Considering the fact that this study was a mixed-method research where the researcher made use of both quantitative and qualitative data, the researcher had an etic role while collecting the quantitative data when participants score sample student exams in standardization trainings and in the official oral examination. During face-to-face standardization sessions, the researcher engaged in full participant role (Patton, 2002) when the participants received general information about the oral exam, exam rules and procedures, and while grading 3 sample graded oral exams because participants asked questions about the things they did not fully grasp and they received feedbacks after grading sample graded student exams. Another reason of being a full participant was the fact that the researcher is a coordinator in the department who is responsible for training instructors for the official examinations of the school. However, while the participants were grading the 10 sample exams, which was used as a quantitative data collection tool, the researcher had a complete observer role where he did not have interaction with the participants and stayed impartial. The researcher's role while training the participants on the online standardization platform was the complete observer as there was no social interaction with the participants while they were receiving their training. The researcher just monitored the participants' training process on the administrative panel of the platform.

Patton (2014) reported in his book that another important feature of the researcher's role was the "revealedness" which is how much participants have been informed about the study. One of the dimensions of the full disclosure where participants know everything about the study from the beginning. The researcher had a full disclosure prior to the start of the study where participants taking part in the study were informed about the objectives of the research and how their inputs were going to be used in the study. Patton also mentioned a possible problem that might emerge from having full disclosure with the participants. The results of the study might be affected negatively as the attitudes of participants might change as they have been given very detailed information about the purposes and procedures of the study. The researcher also spent a high amount of time with the novice instructors taking part in both online and face-to-face standardization sessions in order to make them comfortable during the training sessions and give more honest responses during the semi-structured interviews.

All in all, the role of the researcher had changed a lot during the implementation of the study. The researcher started with an insider role with full participation while giving general information about the oral exam, rules and procedures of the exam, and grading three sample graded exams in face-to-face standardization. Then, in grading ten sample exams, during the online standardization training, and in the official oral examination, the researcher had an etic role with a full observer role. Creswell (2013) expressed that having different roles throughout the implementation of a study is an example of a remarkable qualitative research design.

3.5. Research Design

The researcher employed mixed method research in the study. This design allowed the exploitation of both quantitative and qualitative data, which enabled the researcher gain insights about the various aspects regarding the study (Creswell & Clark, 2007). For the quantitative data of the study, the study utilized the grades of instructors in both standardization groups and official oral examination of the institution they conducted as raters. For the qualitative part of the study, the researcher employed semi-structured interviews in order to gain more insights about the study.

Table 3.2. *Non-equivalent Control Group Post-test-only Quasi-Experimental Design of the Quantitative Part of the Study*

Groups	Standardization Training	Post-Test Measurement
Treatment Group	Instructors were trained on an online platform with regard to the official oral examination of the institution.* They graded 10 sample speaking exams** on an online platform.	Measurement of instructors’ grades in the official oral examination of the institution.
Control Group	Instructors were trained face-to-face with regard to the official oral examination of the institution.* They graded 10 sample speaking exams** on a paper marking sheet.	Measurement of instructors’ grades in the official oral examination of the institution.

* Instructors in both standardization groups received the same training content.
** Instructors in both standardization groups graded the same 10 sample speaking exams.

For the quantitative part of the study, quasi-experimental non-equivalent control group post-test-only design (See Table 3.2.) was resorted since the independent variable of the study (the participants) was manipulated, and the participants in the study were not randomly assigned to their groups, and unlike an experimental research, there was a pre-existing factor, which was the assignment of participants into groups based on their experience in the school's oral examination. Furthermore, the researcher employed a non-equivalent control group post-test-only design because the aim of the researcher was to create two standardization groups (online and face-to-face) where both experienced and novice instructors are assigned in equal numbers. This way, both control group and treatment group could meet the requirement of having particular pre-existing factor, which is having similar backgrounds such as experience as raters in the official oral examination of the school (Cook & Campbell, 1979).

For this study, it was impossible to create random groups of participants since the objective of the researcher was to include instructors with similar backgrounds such as work experience in the school's oral exam. In other words, the researcher wanted to create two standardization groups (online and face-to-face) in which instructors were assigned according to their work experience in the school and experience in the school's official oral examination. As also stated by Cook and Campbell (1979), the quasi-experimental research design could be used to demonstrate whether an educational treatment would prove effective or not in a particular area of study. Remler and Van Ryzin (2011) argued that quasi-experimental studies could be of great significance since they are more practical to conduct, having less ethical limitations, being easier to generalize, relevant regarding the policy of schools, and much easier to conduct in particular institutions where researchers could design them according to their programs and curriculum. On the other hand, Gribbons and Herman (1997) discussed that non-equivalent control group post-test-only design is prone to a significant problem called "selection difference". This problem is the differences between samples, which cannot be controlled or manipulated by the researcher. For instance, the researcher is not able to manipulate the differences such as motivation levels, preparation, being better users of technological devices, etc. Therefore, these selection differences might affect the overall result of the study.

Overall, in quasi-experimental research, employing non-equivalent control group post-test-only design, it could be demonstrated whether measuring a dependent variable (the official oral examination in this study) after treatment in the treatment group (online standardization) and giving the same content to control group (face-to-face standardization) without treatment would yield significant differences in terms of the effectiveness of the treatment (online standardization platform). The independent variable in the study was the training on both standardization groups, which included the general information about the official oral examination such as exam procedures, exam rules, and learning how to grade speaking exams by grading 3 sample graded speaking exams. On the other hand, the dependent variables of the study were the grades given for 10 sample speaking exams in both standardization groups, and the grades given by the instructors in the official oral examination of the school.

The qualitative part of the study aimed to gather broader insights about the online standardization platform which was used to train instructors in the treatment group. Therefore, semi-structured interviews were conducted in order to obtain subjective answers from the participants. This way, it enabled the researcher to explore deeper accounts of the participants' experiences on the online standardization platform. The study by Choak (2013) also expressed that during the implementation of the interview schedule, other themes related to the interview questions are likely to emerge. Thus, it might provide the researcher with additional valuable information and themes that have not been anticipated by the researcher before. To analyze the inputs from the participants, thematic analysis was used. Braun and Clarke (2006) defined thematic analysis as the identification of themes and patterns extracted from a set of data. During the transcription process, a researcher is able to capture significant themes or viewpoints related to research questions and the objective of the study itself. To conclude, thematic analysis would yield to valuable themes and information for the researcher in order to be able to reflect solid outcomes from the dataset.

Table 3.3. *Summary of the Research Design of the Study*

Quantitative Part (Non-equivalent Control Group Post-Test-Only Quasi-Experimental Design	Qualitative Part (Semi-structured Interviews)
Independent variable: General information about the oral exam such as exam procedures, rules, and learning how to grade sample oral exams on online standardization platform and in face-to-face standardization	Semi-structured interviews: To explore broader insights about the online standardization platform
Dependent variable: Participants' grades of 10 sample speaking exams (marking sheets in face-to-face standardization and grading logs on online standardization platform	RQ2: What are the views of the participants on the standardization trainings they received before an official oral examination?
RQ1: How effective is an online standardization platform specifically designed to train teachers that will administer oral exams as raters?	a. What are the views of the participants about an online standardization platform specifically designed to train instructors that will administer oral exams as raters?
a. Do raters trained on an online standardization platform score oral exams consistently within their group?	b. What are the views of the participants about a face-to-face standardization used to train instructors that will administer oral exams as raters?
b. Do raters trained in a face-to-face standardization training score oral exams consistently within their group?	
c. Is there a significant difference between the scores of oral exams given by the raters trained on an online standardization platform and the ones trained in a face-to-face standardization training?	

Dependent variable: Grades of all participants in the actual official oral examination (Approximately 30 speaking exams)	
d. Do raters trained in a face-to-face standardization session and the ones trained on an online standardization platform apply oral exam criteria consistently in an actual oral examination?	

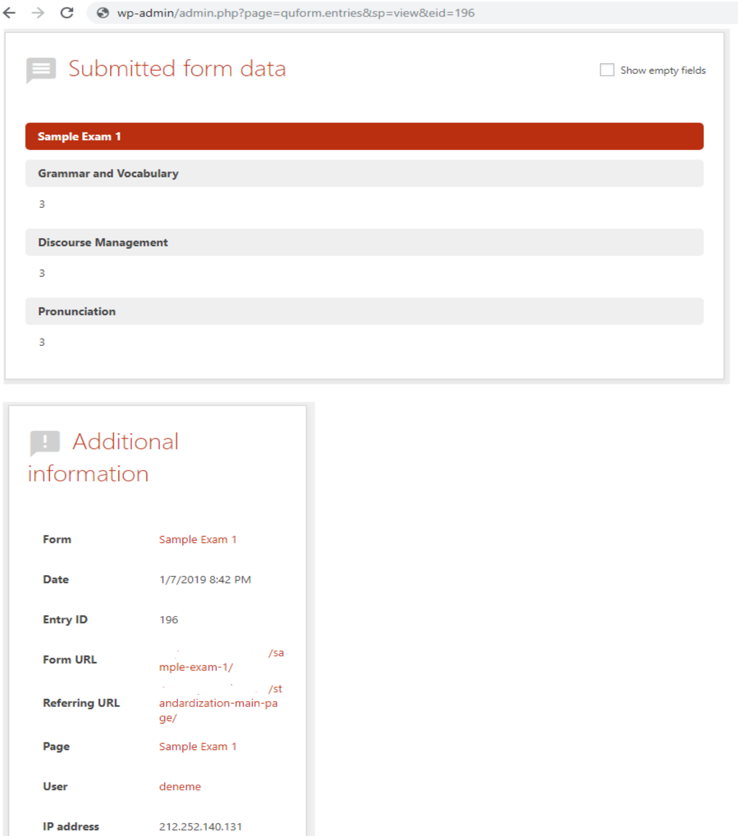
3.6. Data Collection Tools

The data collection tools adopted in the study were grading logs submitted by the instructors trained on online standardization platform, paper grading sheets used by the instructors trained in face-to-face standardization, paper grading sheets used by all instructors trained on both online standardization platform and in face-to-face standardization in the official oral examination of the school, and the semi-structured interviews conducted with the instructors who were trained in both standardization mediums.

As shown in the Figure 3.1., grading logs submitted by the instructors trained on the online standardization platform, paper grading sheets used by the instructors trained in face-to-face standardization (See Figure 3.2.), paper grading sheets (See Figure 3.2.) used by all instructors trained in both face-to-face and online standardization platforms in the actual official oral examination were used in the quantitative part of the study namely quasi-experimental part. And the semi-structured interview was used in the qualitative part of the study. (See Appendix A)

3.6.1. Grade Entries on the Online Standardization Platform

The second data collection tool for the quantitative part of the study was the grade entries on the online standardization platform. The grade entries contained the necessary information such as the particular grades appointed for each part of the criteria (Grammar and Vocabulary, Discourse Management, and Pronunciation), sample exam information (e.g. Sample Exam 1), submission date and time of the entry, the ID of the instructor, and the IP address of the device on which the entry was submitted.



wp-admin/admin.php?page=quform.entries&sp=view&eid=196

Submitted form data

□ Show empty fields

Sample Exam 1

Grammar and Vocabulary

3

Discourse Management

3

Pronunciation

3

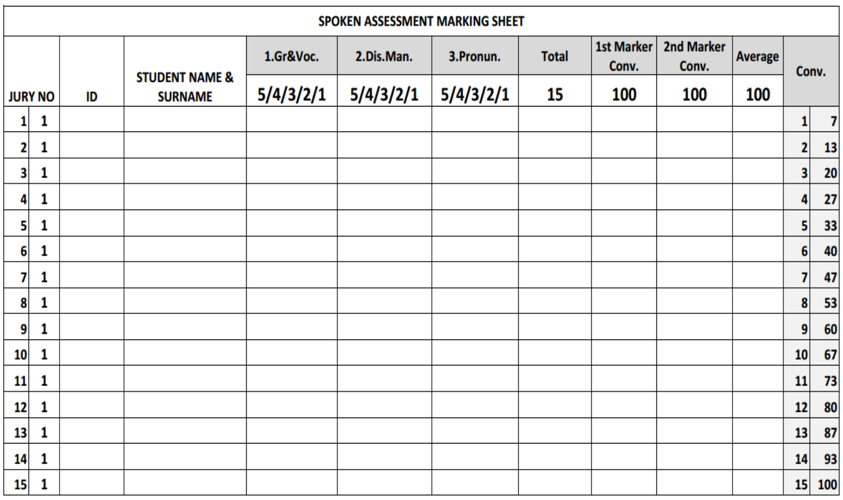
Additional information

Form	Sample Exam 1
Date	1/7/2019 8:42 PM
Entry ID	196
Form URL	/sample-exam-1/
Referring URL	/standardization-main-page/
Page	Sample Exam 1
User	deneme
IP address	212.252.140.131

Figure 3.1. Sample Grade Log Submitted on the Online Standardization Platform

3.6.2. Paper Grading Sheets (Spoken Assessment Marking Sheet)

As the first data collection tool of the study, paper grading sheets were used by the instructors trained in face-to-face standardization, and then these sheets were used by all instructors, who had been trained both in person and online, in the official oral examination of the institution. The researcher analyzed the grading sheets collected in face-to-face standardizations and the ones from the official oral examination for the quantitative part of the study.



SPOKEN ASSESSMENT MARKING SHEET

JURY NO		ID	STUDENT NAME & SURNAME	1.Gr&Voc.	2.Dis.Man.	3.Pronun.	Total	1st Marker Conv.	2nd Marker Conv.	Average	Conv.	
				5/4/3/2/1	5/4/3/2/1	5/4/3/2/1	15	100	100	100		
1	1										1	7
2	1										2	13
3	1										3	20
4	1										4	27
5	1										5	33
6	1										6	40
7	1										7	47
8	1										8	53
9	1										9	60
10	1										10	67
11	1										11	73
12	1										12	80
13	1										13	87
14	1										14	93
15	1										15	100

Figure 3.2. Paper Grading Sheets used in Face-to-face Standardization and in the Official Oral Examination

The Spoken Assessment marking sheet is a printed grading sheet used to grade students in an oral examination. It consists of several columns under which the raters of the exam must provide grades after conducting an oral examination. The sheet includes information about the Jury, which refers to the classroom where particular instructors conduct the exam, ID number of the students, names of the students, the parts of the criteria (Grammar and Vocabulary, Discourse Management, Pronunciation) where instructors grade students (Out of 5 for each part, 15 total) by addressing to the criteria, the total grade of the student (between 0 and 15), the converted grade of the 1st instructor (Out of 100), the converted grade of the 2nd instructor (Out of 100), the average grade of the student (Out of 100), and lastly the conversion table where instructors can convert their grades from (0-15) to the ones (0-100).

3.6.3. Semi-structured Interviews

At the end of the standardization process, the researcher implemented semi-structured interviews with instructors trained both in face-to-face standardization and on the online standardization platform. All instructors were asked the same interview questions (See Appendix A). Although the main focus of the study was to gather accounts of instructors' experiences on the online standardization platform, the researcher had the idea that exploring the experiences of instructors in face-to-face standardization would also help him to discover valuable themes for the results of the study. To illustrate, as instructors trained online and face-to-face were asked the same interview questions, it was likely that they could provide answers that overlap. For instance, while one of the instructors trained online mentioned that online standardization platform is practical as it enables instructors to train anytime and there was the possibility of giving breaks whenever needed, another instructor trained face-to-face expressed that there was not enough time to cover all the training content in one session as there was not a suitable time period due to the schedules of instructors. This example can be seen as a justification why the researcher decided to include interviews with the instructors trained in person so that overlapping answers could be benefited to provide meaningful themes for the results chapter.

Table 3.4. *Summary of the Instruments Used to Gather Data for the Study*

Research Questions	Method	Instruments
RQ1: How effective is an online standardization platform specifically designed to train teachers that will administer oral exams as raters?	Quantitative	Collected scores of the instructors in face-to-face standardization sessions, on online standardization platform, in the official oral examinations of the school
a. Do raters trained on an online standardization platform score oral exams consistently within their group?	Quantitative	Collected scores of the instructors in face-to-face standardization sessions
b. Do raters trained in a face-to-face standardization training score oral exams consistently within their group?	Quantitative	Collected scores of the instructors on online standardization platform

c. Is there a significant difference between the scores of oral exams given by the raters trained on an online standardization platform and the ones trained in a face-to-face standardization training?	Quantitative	Collected scores of the instructors in face-to-face standardization sessions, on online standardization platform
d. Do raters trained on an online standardization platform and the ones trained in a face-to-face standardization session apply oral exam criteria consistently in an actual oral examination?	Quantitative	Collected scores of the instructors in the official oral examinations of the school
RQ2: What are the views of the participants about the standardization trainings they received before an official oral examination?	Qualitative	Semi-structured interviews conducted with participants in both groups
a. What are the views of the participants about an online standardization platform specifically designed to train teachers that will administer oral exams as raters?	Qualitative	Semi-structured Interviews
b. What are the views of the participants about a face-to-face standardization used to train teachers that will administer oral exams as raters?	Qualitative	Semi-structured Interviews

3.7. Implementation Process and Procedures

The implementation process and procedures of the study launched after receiving consent letters from the instructors stating that they participated in the study intentionally, and acquiring the approval of METU Ethics Committee (See Appendix B). Before online and face-to-face standardization sessions commenced on 10

December 2018, the researcher informed all instructors taking part in the study that their inputs throughout the sessions would be used in the study.

Prior to the start of the standardization sessions, the researcher notified all instructors about the standardization sessions via e-mail. For the instructors who would be trained in person, the researcher prepared schedules in which instructors could see when, where and with whom they were going to have the training. These schedules were sent to them via e-mail.

For the instructors who would be trained online, the researcher, also the administrator of the online standardization platform, created login credentials and specified due dates for the completion of the standardization. Then, the researcher notified these instructors that they were going to take a short training on how to use the platform in person. Five-minute meeting schedules were prepared by the researcher and sent to instructor via e-mail. In these five-minute meetings, the researcher trained instructors on how to use the online standardization platform in person in order to make them familiar with the training environment. After every instructor to be trained online was briefed in these meetings, the researcher sent every instructor their login credentials and due dates (10 December 2018 – 17 December 2018) for the completion of the standardization via e-mail.

The study is conducted in three main phases. The first phase included the above-mentioned standardization sessions, the second phase included semi-structured interviews, and the third phase included the official oral examination of the school, which was the post-test of the study (See Table 3.5.).

Table 3.5. *Phases of the Study*

Groups		Phase 1	Phase 2	Phase 3
Online	6 Novice - 6 Experienced Instructors	Online standardization platform (Treatment) General information about the oral exam such as exam procedures, rules, and learning how to grade sample oral exams, scoring 3 sample graded exams, scoring 10 student exams from the previous years	Semi-structured Interviews	Spoken Assessment (Official Oral Examination) Collected scores of the instructors
Face-to-face	6 Novice - 6 Experienced Instructors	Face-to-face standardization sessions (No treatment) General information about the oral exam such as exam procedures, rules, and learning how to grade sample oral exams, scoring 3 sample graded exams, scoring 10 student exams from the previous years	Semi-structured Interviews	Spoken Assessment (Official Oral Examination) Collected scores of the instructors

3.7.1. Phase 1: Online Standardization Platform

In online standardization session of the first phase, the researcher sent the URL of the standardization platform and due dates for the completion of the standardization to the instructors via e-mail.

Instructors were told that they could give breaks anytime they wish ensuring that they had submitted their scores if they had been scoring any at that time. If they had not been scoring any exams, instructors were free to give breaks at any time intervals and come back again any time they wanted.

As it was an online standardization platform, the instructors could reach the platform anywhere as long as they have Internet connection. However, completing the session in a quiet place was advised as unwanted noise or distractions could hinder their performance on the scoring. The platform could also be benefited by smart phones and tablets as it was a fully responsive platform.

Contrary to face-to-face standardization session, instructors were not given any verbal explanations, apart from the informative briefings on how to use the platform, since all the necessary information was provided on the online standardization platform. The platform had texts, pictures and visual aids explaining each component of the oral exam such as the parts of the exam, rules, exam criteria and the grading sheet of the exam.

All instructors taking part on the online standardization platform finished their trainings in a week. As it could be seen from the admin panel, some of them finished the standardization in one day, some of them preferred to give two or three-day breaks between the sample exams. As opposed to the face-to-face standardization session, the instructors could receive feedbacks right after they had finished scoring an exam because they were re-directed to the feedback page as soon as they submitted their scores of any particular exam on the platform.

Overview of the Online Standardization Platform

In June, 2018 the researcher created an online platform where instructors can refresh their knowledge about the oral exams of the school and score sample exams for the actual oral examinations in the institution. The researcher designed an online standardization platform where instructors could be provided with the same contents and materials as in a conventional face-to-face standardization session. The idea of creating an online standardization platform resulted from a needs analysis conducted in the year of 2016. That year, the institution had problems in carrying out standardization sessions because the course schedules of teachers were loaded and the school coordinators had difficulties in appointing suitable time periods in which all teachers could show up and take the standardization session. As a result, the institution decided to conduct several standardization sessions by gathering the available teachers at a time. This resulted in the need for a solution that can help both teachers and coordinators in the institution. As part of professional development activities, the researcher came up with the idea that an online standardization platform to train teachers for the oral examinations in the school could solve the institution's main problem of gathering all teachers for the standardization.

Early Design of the Online Standardization Platform

The online standardization platform was first designed and tried in 2016 as part of a professional development study, and it was presented to the instructors and guests in an English Language Teaching event at the department of foreign languages of the researcher's institution. The first online standardization platform was created as a Google Form document in which instructors were trained in three sessions. Instructors first started the standardization by providing their credentials. In the first session, they were informed about the exam procedures and exam rules. The written information was uploaded as pictures onto the form as Google did not allow to change font color and size of the texts at that time. Therefore, the researcher had to take screenshots from the institution's documents in order to upload them onto the form in a way that instructors could read without problems. However, because of the layout of Google Forms, it was really problematic to insert so many pictures with the same width and height onto the form. In the 2nd and 3rd session of the standardization, the instructors were asked to read the criteria of the exam and grade sample speaking exams, which were extracted from real oral examinations conducted in the past. The parts of the exam and the audio of the students were merged into a video file so that instructors could see what students were talking about and which pictures they were describing throughout the video. The videos were uploaded to YouTube and embedded onto the form below the criteria of the exam (See Figure 3.5., Appendix C). This situation also brought about problems related to navigation and the quality of grading. As the layout did not allow to put the criteria and the video next to each other, the instructors had to scroll down and up while watching and listening to the sample student exam, which created an extraneous overload and distraction for them. Several instructors using the platform mentioned that they had difficulty in grading the students as they constantly had to scroll up to read the criteria and scroll down to continue to the video. Instructors could stop the video, go backwards and forwards in the video when they had to think carefully on what grade to give. This time, they had to go back to the exam criteria to be sure about the grade. To do this, scrolling up and down on the platform was inevitable, about which several instructors complained after completion. The platform was also presented in another ELT event held at a foundation university in Ankara in the following year, and it received a lot of positive feedbacks from the participants coming from different institutions in Ankara. That paved the way to a better and thoroughly thought version of the platform.

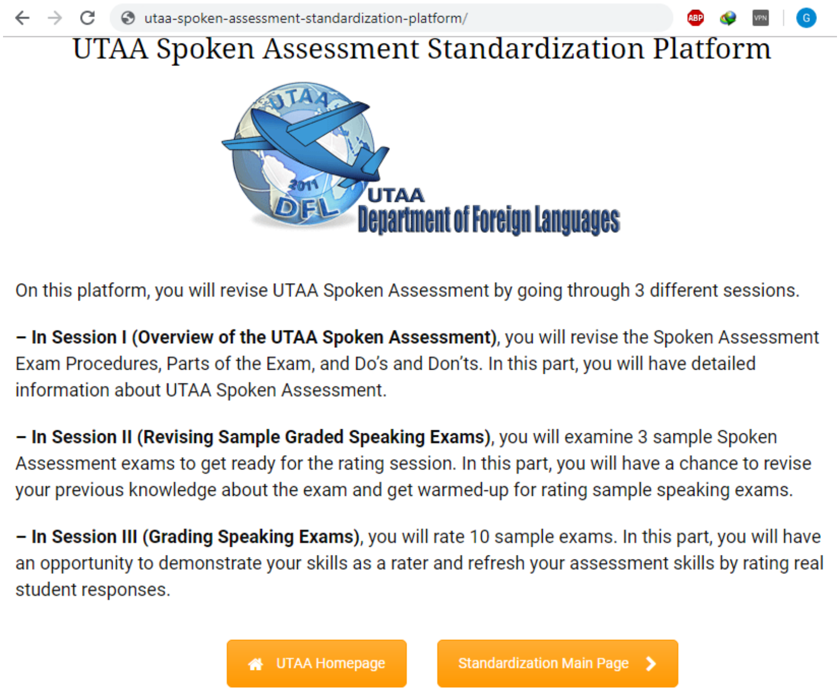
The early version of the platform was created as a professional development tool and it was not used by the participants who took part in this study.

The Online Standardization Platform (Used in the study)

After developing the first version of the online standardization platform on Google Forms, the researcher created a WordPress website which enabled him to manipulate the layout of the website by using HTML and CSS. In this way, the researcher was able to turn his ideas into reality on an online standardization platform.

Unlike a regular website, the platform did not have any main menus at the top of the page and side menus on the left and the right of the page. This way, the platform could be provided to instructors as a full-page platform, which eliminated extraneous overload. It also helped to put written information with bigger font sizes and different ranges of color.

To reach the standardization platform, the instructors had to have login credentials provided by the admin, who was also the researcher of the study. Otherwise, they could not be able to enter the platform although they had the exact link of the platform.


utaa-spoken-assessment-standardization-platform/
UTAA
2011
DFL
UTAA
Department of Foreign Languages
UTAA Homepage
Standardization Main Page

UTAA Spoken Assessment Standardization Platform

On this platform, you will revise UTAA Spoken Assessment by going through 3 different sessions.

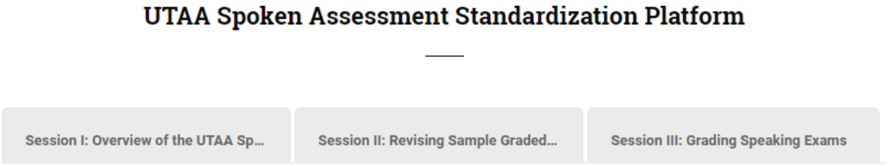
– In Session I (Overview of the UTAA Spoken Assessment), you will revise the Spoken Assessment Exam Procedures, Parts of the Exam, and Do's and Don'ts. In this part, you will have detailed information about UTAA Spoken Assessment.

– In Session II (Revising Sample Graded Speaking Exams), you will examine 3 sample Spoken Assessment exams to get ready for the rating session. In this part, you will have a chance to revise your previous knowledge about the exam and get warmed-up for rating sample speaking exams.

– In Session III (Grading Speaking Exams), you will rate 10 sample exams. In this part, you will have an opportunity to demonstrate your skills as a rater and refresh your assessment skills by rating real student responses.

Figure 3.3. Landing Page of the Online Standardization Platform Used in the Study

The main page (See Figure 3.4.) of the standardization platform has tabbed menu that has three sections on it. When the page loads the first time, instructors see the three sections as closed, which was decided by the researcher on the admin panel.


Session I: Overview of the UTAA Sp...
Session II: Revising Sample Graded...
Session III: Grading Speaking Exams

UTAA Spoken Assessment Standardization Platform

Figure 3.4. Main Page of the Platform with Tabbed Menus

When clicked on, a section on the page collapse, and when another section is clicked the previous section closes and the recently clicked section collapses on the page. As in the early version, the current version of the platform provided instructors with general information about the exam, the exam procedures, and the rules of the exam as embedded video files in the 1st part of the standardization. The 1st section of the standardization is composed of accordion menus (See Figure 3.5.) which collapse upon clicking and close when clicked for the second time (See Figure 3.6.) Unlike the tabbed menu, the items in the accordion menu will not collapse unless clicked on a second time.

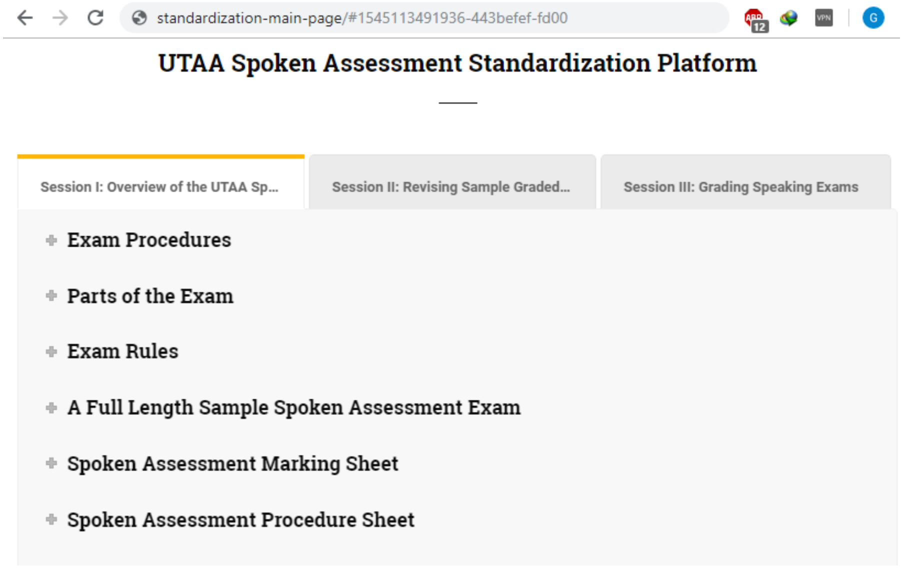
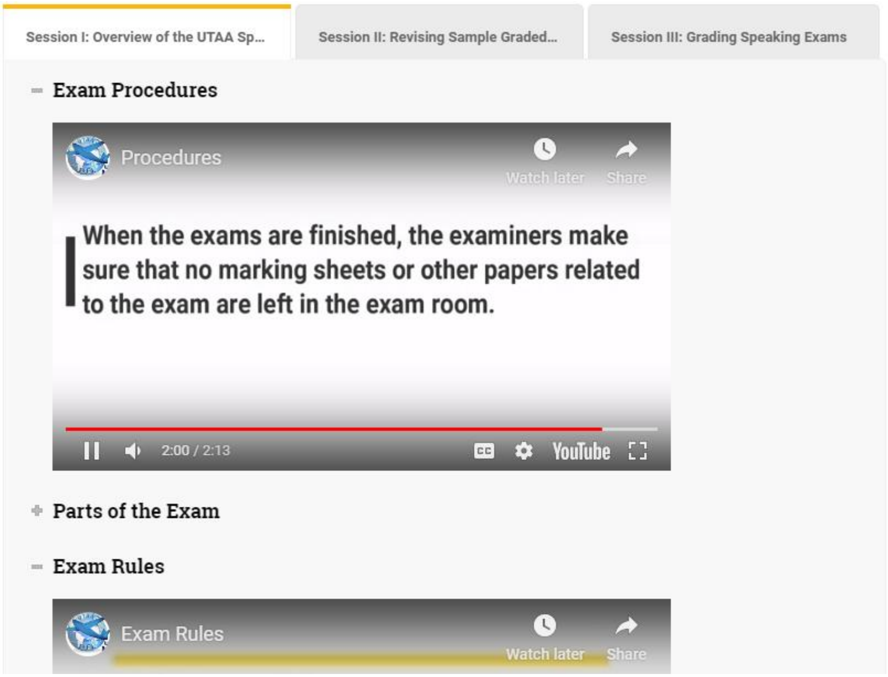

standardization-main-page/#1545113491936-443befef-fd00
UTAA Spoken Assessment Standardization Platform
Session I: Overview of the UTAA Sp...
Session II: Revising Sample Graded...
Session III: Grading Speaking Exams
Exam Procedures
Parts of the Exam
Exam Rules
A Full Length Sample Spoken Assessment Exam
Spoken Assessment Marking Sheet
Spoken Assessment Procedure Sheet

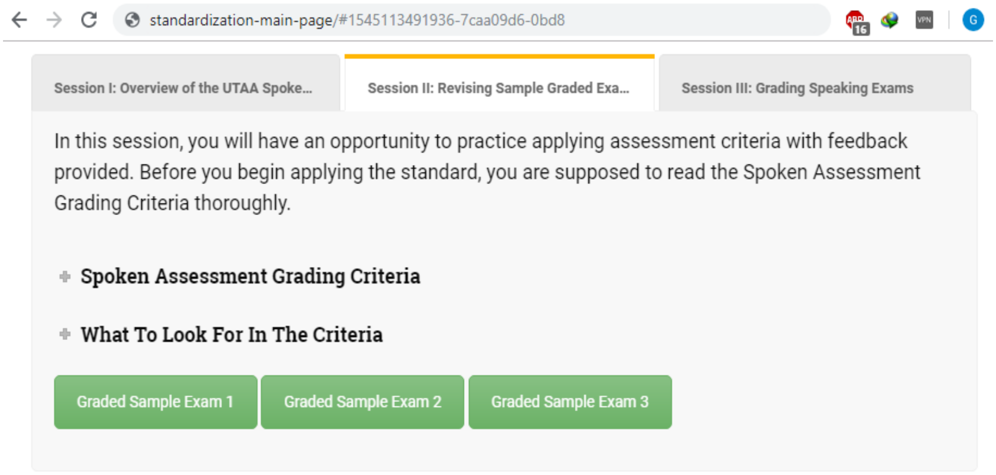
Figure 3.5. Accordion Menus of the Platform



When the exams are finished, the examiners make sure that no marking sheets or other papers related to the exam are left in the exam room.

Figure 3.6. Collapsed and Closed Accordion Menus on the Platform

The 2nd part of the standardization is the Revising Sample Graded Speaking Exams session (See Figure 3.7.) where instructors can have a look at the exam criteria in detail and learn what to expect from students and how to grade them by addressing to the parts of the criteria (Grammar and Vocabulary, Discourse Management, Pronunciation).



standardization-main-page/#1545113491936-7caa09d6-0bd8

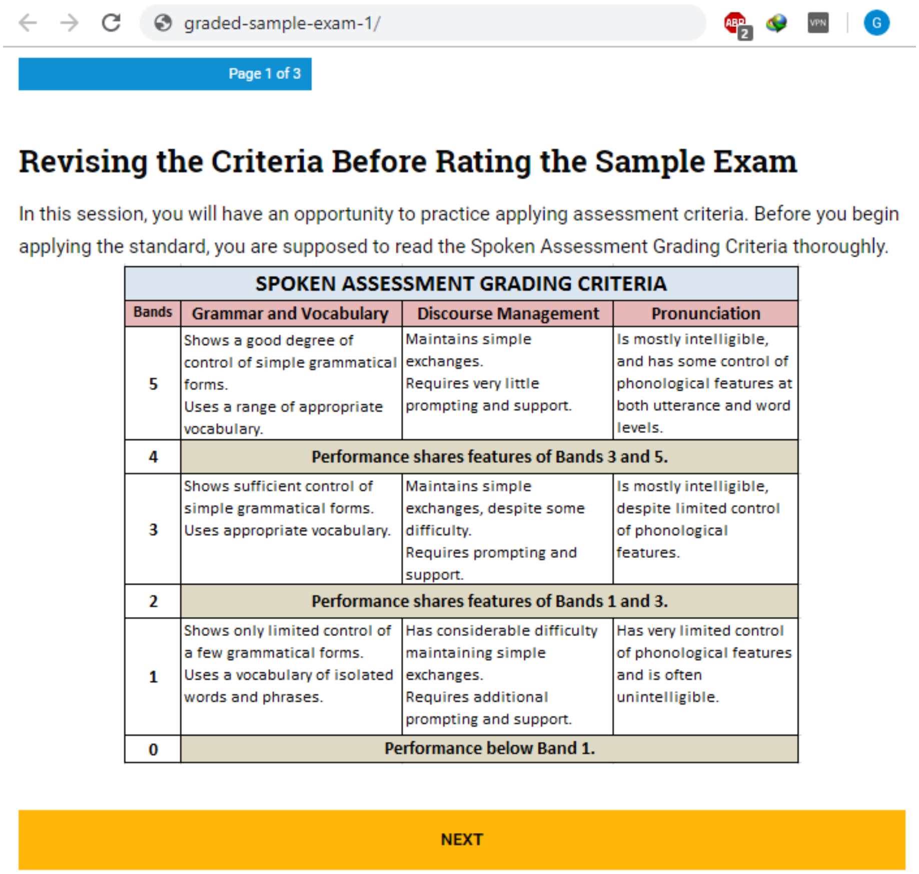
Session I: Overview of the UTAA Spoke...
Session II: Revising Sample Graded Exa...
Session III: Grading Speaking Exams

In this session, you will have an opportunity to practice applying assessment criteria with feedback provided. Before you begin applying the standard, you are supposed to read the Spoken Assessment Grading Criteria thoroughly.

Spoken Assessment Grading Criteria
What To Look For In The Criteria
Graded Sample Exam 1
Graded Sample Exam 2
Graded Sample Exam 3

Figure 3.7. Second Tabbed Menu of the Platform (Revising Sample Graded Exams)

This session is of great importance for the instructors as it enables them to understand the parts of the criteria thoroughly so that they would not have any questions in their minds while grading sample speaking exams and real oral examinations in the future. Upon looking at the criteria and receiving detailed information about the parts of the oral examination criteria (See Figure 3.8.), the instructors can move on to grading sample exams by clicking on the green menu buttons. The buttons were coded as “_blank” so that they open in another tab on the browser. This way, instructors do not have to leave the last page they have been on. When they are finished with the grading of a sample exam, they can close the tab and continue from where they left on the page.



Revising the Criteria Before Rating the Sample Exam

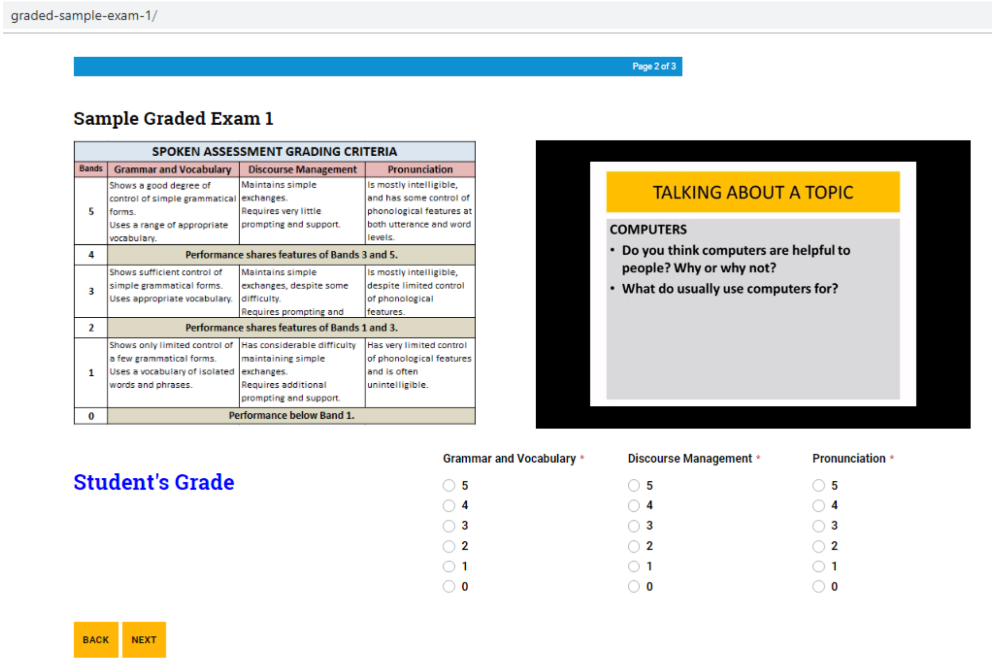
In this session, you will have an opportunity to practice applying assessment criteria. Before you begin applying the standard, you are supposed to read the Spoken Assessment Grading Criteria thoroughly.

SPOKEN ASSESSMENT GRADING CRITERIA

Bands	Grammar and Vocabulary	Discourse Management	Pronunciation
5	Shows a good degree of control of simple grammatical forms. Uses a range of appropriate vocabulary.	Maintains simple exchanges. Requires very little prompting and support.	Is mostly intelligible, and has some control of phonological features at both utterance and word levels.
4	Performance shares features of Bands 3 and 5.		
3	Shows sufficient control of simple grammatical forms. Uses appropriate vocabulary.	Maintains simple exchanges, despite some difficulty. Requires prompting and support.	Is mostly intelligible, despite limited control of phonological features.
2	Performance shares features of Bands 1 and 3.		
1	Shows only limited control of a few grammatical forms. Uses a vocabulary of isolated words and phrases.	Has considerable difficulty maintaining simple exchanges. Requires additional prompting and support.	Has very limited control of phonological features and is often unintelligible.
0	Performance below Band 1.		

Figure 3.8. Grading Criteria in the Revising Sample Graded Exams Session

When a sample graded exam page is opened on the browser, the instructors can see a progress bar at the top of the page, which informs them how many more pages left before the end of that section and provides comfort of knowing how long the section is. In this session, the instructors begin by being required to read the exam criteria in detail before moving on the next page. The idea here is to make instructors familiar with the exam criteria as much as possible when applicable.



graded-sample-exam-1/

Sample Graded Exam 1

SPOKEN ASSESSMENT GRADING CRITERIA			
Bands	Grammar and Vocabulary	Discourse Management	Pronunciation
5	Shows a good degree of control of simple grammatical forms. Uses a range of appropriate vocabulary.	Maintains simple exchanges. Requires very little prompting and support.	Is mostly intelligible, and has some control of phonological features at both utterance and word levels.
4	Performance shares features of Bands 3 and 5.		
3	Shows sufficient control of simple grammatical forms. Uses appropriate vocabulary.	Maintains simple exchanges, despite some difficulty. Requires prompting and	Is mostly intelligible, despite limited control of phonological features.
2	Performance shares features of Bands 1 and 3.		
1	Shows only limited control of a few grammatical forms. Uses a vocabulary of isolated words and phrases.	Has considerable difficulty maintaining simple exchanges. Requires additional prompting and support.	Has very limited control of phonological features and is often unintelligible.
0	Performance below Band 1.		

TALKING ABOUT A TOPIC

COMPUTERS

- Do you think computers are helpful to people? Why or why not?
- What do usually use computers for?

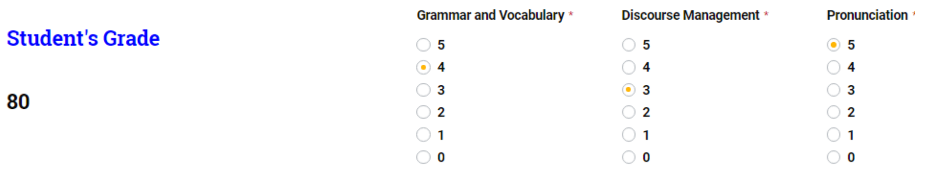
Student's Grade

Grammar and Vocabulary *	Discourse Management *	Pronunciation *
○ 5	○ 5	○ 5
○ 4	○ 4	○ 4
○ 3	○ 3	○ 3
○ 2	○ 2	○ 2
○ 1	○ 1	○ 1
○ 0	○ 0	○ 0

Figure 3.9. The Grading Page of the Platform (Same for all Exams on the Platform)

As it can be seen in Figure 3.9., on the next page, instructors can see the exam criteria again along with the sample student exam prepared as a video file. The exam criteria and the student exam video are placed next to each other so that instructors can have a look at the criteria while watching and listening to the sample oral exam at the same time without having to scroll down or up, looking at another page, or taking a look at a printed exam criteria in front of them, which helps decrease the cognitive overload by addressing to the instructional design theories mentioned in the literature review.

Below the criteria and the student exam video, the instructors can see the grading system where they can easily grade students by clicking on one of the radio buttons. The grading system does not allow multiple responses for a particular column. That is, instructors cannot choose more than one value (0-5) in the radio button list under each part of the criteria. Also, the page does not allow instructors to continue unless they appoint a grade value (0-5) for each of the three parts of the exam criteria.



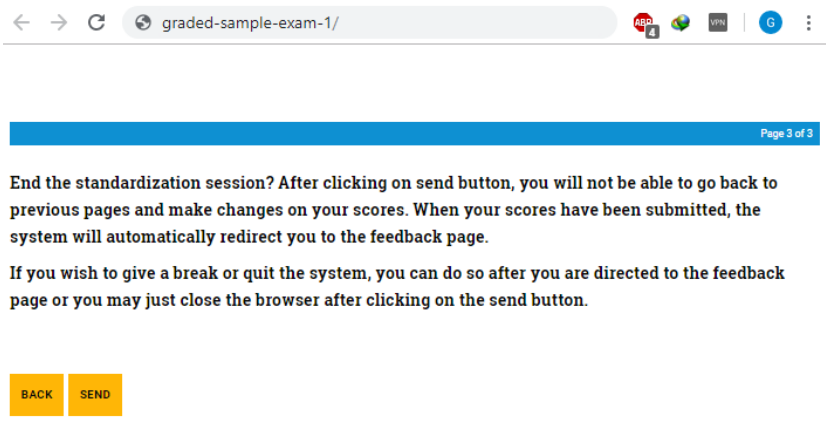
Student's Grade

80

Grammar and Vocabulary *	Discourse Management *	Pronunciation *
○ 5	○ 5	● 5
● 4	○ 4	○ 4
○ 3	● 3	○ 3
○ 2	○ 2	○ 2
○ 1	○ 1	○ 1
○ 0	○ 0	○ 0

Figure 3.10. Automatic Conversion of Grades on the Platform Figure 3.10.

The grading system on the platform is coded in a way that grades are projected on the screen depending on the values of Grammar and Vocabulary, Discourse Management and Pronunciation. Conditional Logic on the platform enables participants to have a more reliable score as it does not allow any errors in converting the values into a score out of hundred like in actual examinations. After providing grades for each of the exam criteria, the system automatically projects the student's grade out of hundred on the page as seen in Figure 3.10.



graded-sample-exam-1/

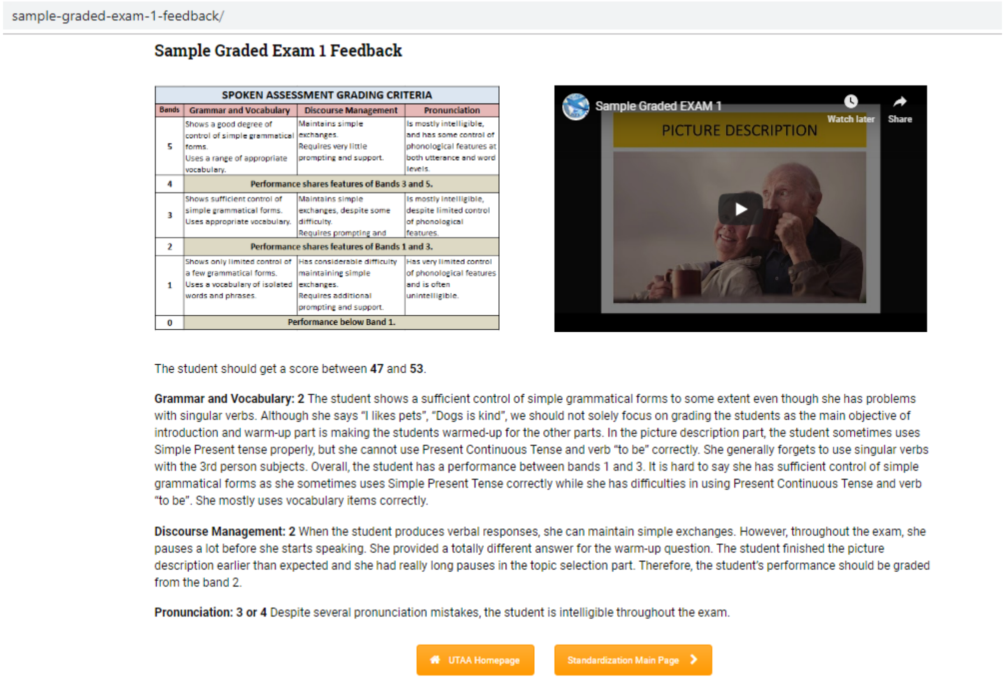
Page 3 of 3

End the standardization session? After clicking on send button, you will not be able to go back to previous pages and make changes on your scores. When your scores have been submitted, the system will automatically redirect you to the feedback page.

If you wish to give a break or quit the system, you can do so after you are directed to the feedback page or you may just close the browser after clicking on the send button.

Figure 3.11. Last Page of the Grading Exams Sessions (Same for all Exams on the Platform)

After grading the student, the instructors can navigate to the previous page or the next below from the navigation buttons at the bottom of the page. After clicking next, the last page of the grading sample student exam session loads. On this page, instructors are informed that they will not be able to make changes on their grades if they click on "send" button. This way, they are given the chance to think about the grades they have given and make changes on them if necessary. If instructors are ready to submit their grades, they can click on "send" button, which will direct them to the feedback page of the exam they have just graded (See Figure 3.11.).



sample-graded-exam-1-feedback/

Sample Graded Exam 1 Feedback

SPOKEN ASSESSMENT GRADING CRITERIA			
Bands	Grammar and Vocabulary	Discourse Management	Pronunciation
5	Shows a good degree of control of simple grammatical forms. Uses a range of appropriate vocabulary.	Maintains simple exchanges. Requires very little prompting and support.	Is mostly intelligible, and has some control of phonological features at both utterance and word levels.
4	Performance shares features of Bands 3 and 5.		
3	Shows sufficient control of simple grammatical forms. Uses appropriate vocabulary.	Maintains simple exchanges, despite some difficulty. Requires prompting and	Is mostly intelligible, despite limited control of phonological features.
2	Performance shares features of Bands 1 and 3.		
1	Shows only limited control of a few grammatical forms. Uses a vocabulary of isolated words and phrases.	Has considerable difficulty maintaining simple exchanges. Requires additional prompting and support.	Has very limited control of phonological features and is often unintelligible.
0	Performance below Band 1.		

The student should get a score between **47** and **53**.

Grammar and Vocabulary: 2 The student shows a sufficient control of simple grammatical forms to some extent even though she has problems with singular verbs. Although she says "I likes pets", "Dogs is kind", we should not solely focus on grading the students as the main objective of introduction and warm-up part is making the students warmed-up for the other parts. In the picture description part, the student sometimes uses Simple Present tense properly, but she cannot use Present Continuous Tense and verb "to be" correctly. She generally forgets to use singular verbs with the 3rd person subjects. Overall, the student has a performance between bands 1 and 3. It is hard to say she has sufficient control of simple grammatical forms as she sometimes uses Simple Present Tense correctly while she has difficulties in using Present Continuous Tense and verb "to be". She mostly uses vocabulary items correctly.

Discourse Management: 2 When the student produces verbal responses, she can maintain simple exchanges. However, throughout the exam, she pauses a lot before she starts speaking. She provided a totally different answer for the warm-up question. The student finished the picture description earlier than expected and she had really long pauses in the topic selection part. Therefore, the student's performance should be graded from the band 2.

Pronunciation: 3 or 4 Despite several pronunciation mistakes, the student is intelligible throughout the exam.

Figure 3.12. Feedback Page on the Platform (Same Layout for all Exams on the Platform)

Feedback pages (See Figure 3.12.) are provided for all sample exams on the platform. In the 2nd session, there are 3 sample exams and in the 3rd session of the standardization, there are 10 sample exams. For each of these exams, the researcher created a feedback page, where instructors can compare their grades with the actual grades the students should receive. Each feedback page provides detailed explanations about why students should get a particular grade by referring to the exact mistakes they have done in their exams. All mistakes made by the students are provided on feedback pages so that instructors might not miss them and understand why they gave the correct grade or why they should have given a different grade. In addition to these, instructors can watch and listen to the student exam again on the feedback page as the criteria and the video are provided on the page. However, they are not able to grade it this time.

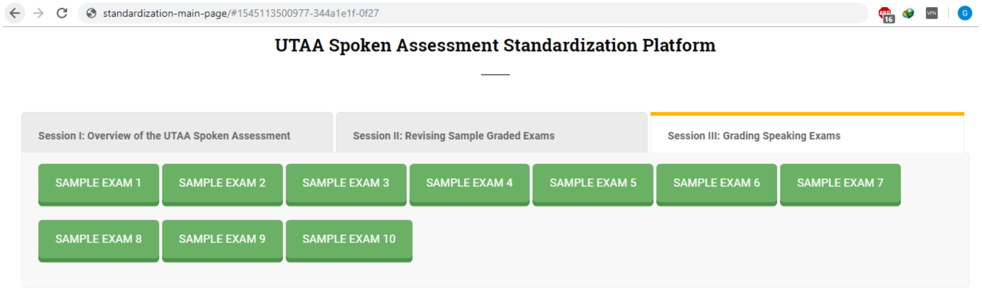
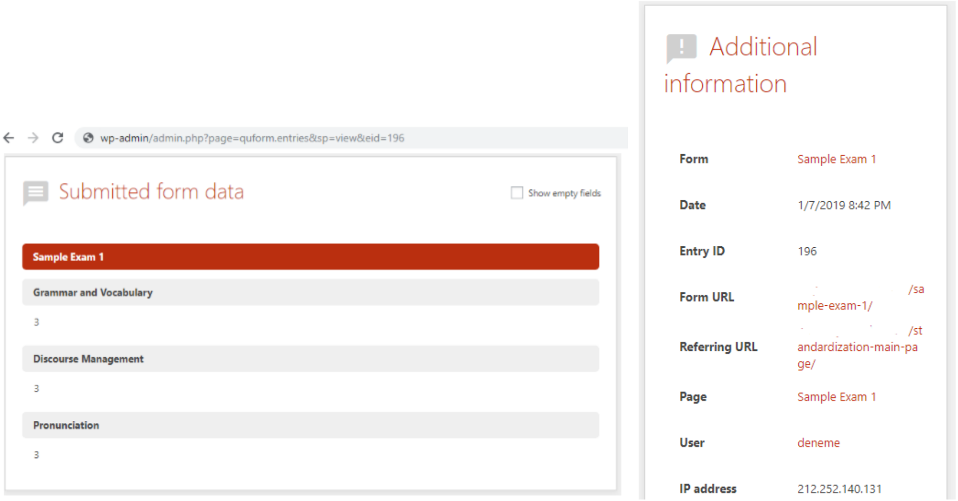

standardization-main-page/#1545113500977-344a1e1f-0f27
UTAA Spoken Assessment Standardization Platform
Session I: Overview of the UTAA Spoken Assessment
Session II: Revising Sample Graded Exams
Session III: Grading Speaking Exams
SAMPLE EXAM 1
SAMPLE EXAM 2
SAMPLE EXAM 3
SAMPLE EXAM 4
SAMPLE EXAM 5
SAMPLE EXAM 6
SAMPLE EXAM 7
SAMPLE EXAM 8
SAMPLE EXAM 9
SAMPLE EXAM 10

Figure 3.13. Third Tabbed Menu of the Platform (Grading Sample Speaking Exams)

The 3rd session of the standardization is Grading Sample Exams part (See Figure 3.13.) where instructors grade 10 sample student exams. The idea of providing instructors with 10 sample speaking exams is that these exams differed in terms of students' speaking skills. All exams in this session had at least three samples of a low achiever, average, and a high achiever student ordered randomly.



Sample Exam 1

Grammar and Vocabulary	3
Discourse Management	3
Pronunciation	3

Form	Sample Exam 1
Date	1/7/2019 8:42 PM
Entry ID	196
Form URL	/sample-exam-1/
Referring URL	/standardization-main-page/
Page	Sample Exam 1
User	deneme
IP address	212.252.140.131

Figure 3.14. Submitted Exam Data of the Trainees on the Platform

When instructors submit their scores, they are stored in the admin panel so that the researcher could see which teacher scored which particular student and their grades for those particular students.

3.7.2. Phase 1: Face-to-face Standardization

The researcher conducted face-to-face standardization in 8 sessions in two weeks (See Table 3.6.) as it was not possible to gather all instructors in a room at the same time.

As instructors' schedules were busy, the researcher first had to specify a suitable date and time period for particular instructors.

Table 3.6. *Meetings with the Instructors Trained in Face-to-face Standardization*

Groups	1st Week	2nd Week
1st Group	4 instructors	4 instructors
2nd Group	3 instructors	3 instructors
3rd Group	3 instructors	3 instructors
4th Group	2 instructors	2 instructors
Total	12 instructors	12 instructors

Therefore, face-to-face standardization could be completed in two weeks where all instructors received half of the training in the first week and received the rest in the second week. In total, the researcher had to specify 4 groups for the standardization because of the limitations mentioned above. The researcher met these 4 groups twice (once for a week) in order to complete the standardization. In total, there were 8 standardization meetings for face-to-face standardization.

First Week of Face-to-face Standardization

At the beginning of the session, the researcher welcomed the instructors and reminded them that their inputs would be used in the study once again and thanked them for their contributions. Then, the researcher informed instructors that the session would be video-recorded and started recording the session. The camera was placed behind the instructors so that they would not be distracted by it during the session. Later, the researcher stated that they were going to analyze the oral exam criteria and exam marking sheet. The parts of the oral exam were explained and instructors were provided information about what to do and what not to do during the administration of the exam. (General information about the oral exam, exam procedures and exam rules)

Each participant in the face-to-face standardization session received a copy of the criteria and marking sheet used in institution’s official oral examination. Then, they were asked to read the criteria and marking sheet thoroughly before moving on to scoring 3 sample graded exams.

After making sure that everybody was ready to start scoring graded oral exams, the researcher used audio recordings of students’ oral exams prepared as a video file and played them in the session room. Teachers listened to the audio recordings as if they were listening to a student during the official oral exam, and they scored 3 sample graded exams on marking sheets by addressing to the criteria. The exam audios of the students were merged into a video file so that instructors could see what students were talking about and which pictures they were describing throughout the video. On the online standardization platform, the same videos were used as embedded YouTube videos.

Having finished the scoring of sample exams, the researcher explained the grades of each exam in detail by referring to the parts of the criteria. All three exams were selected from different levels of proficiency (1 low score, 1 average score and 1 high score) so that teachers could be exposed to different parts in the criteria.

The researcher then asked whether there were any differences between the teachers’ scores and the actual scores of the students. Teachers were provided with vital information about each sample exam’s scores such as why the student should get a higher grade and vice versa.

Lastly, after having settled on all the scores of the 3 sample graded exams, the researcher went on the next part where instructors were supposed to score 10 sample speaking exams on their own. However, due to the time limitations, instructors could only score 2 sample speaking exams and gave feedbacks on why those exams should have received particular scores by referring to the exam criteria. Therefore, the first meeting of face-to-face standardization had to end. The researcher then kindly asked teachers to hand in their marking sheets and informed instructors that they would meet next week for the rest of the sample exams.

Second Week of Face-to-face Standardization

At the beginning of the session, the researcher welcomed the instructors and informed them that the session would be video-recorded as in the previous week and started recording the session. Like in the previous week, the camera was placed behind the instructors so that they would not be distracted by it during the session.

Later, the researcher reminded instructors that they had covered general information about the exam, exam procedures, exam rules about the exam, three sample graded speaking exams, and two sample speaking exams. Then, the researcher asked whether they had any questions so that he could answer before moving onto scoring the remaining 8 sample speaking exams. When instructors had no questions on their minds and were ready to start, the researcher opened the file where sample speaking exams were stored and started playing the remaining exams one by one. This time, between sample exams, there were no breaks and feedback sessions because of the time limitations. In school's conventional face-to-face standardizations in the past, there had usually been maximum 5 sample exams which were thoroughly analyzed by the coordinators and instructors taking part in that session. However, as there were 8 sample exams in this session and the time was limited, feedbacks were given after the standardization ended as printed documents. Instructors received the same feedbacks as the ones trained online, but the feedbacks were printed documents instead of being online. When instructors finished scoring a total of 8 sample speaking exams, the researcher told them that their standardization was over, and thanked them for their precious contributions to the study.

3.7.3. Phase 2: Semi-structure Interviews

The researcher conducted 22 interviews (12 online, 10 face-to-face) with the participants of the study. Although the instructors who could not take part in the interview gave their consent to be interviewed by the researcher, they could not take the interview due to personal reasons. The duration of the interviews varied between three to five minutes (See Table 3.7.)

Table 3.7. *Durations of the Interviews with the Instructors in Both Mediums*

Face-to-face	Duration (Minute)	Online	Duration (Minute)
Instructor 1	05:23	Instructor 1	05:40
Instructor 2	04:41	Instructor 2	05:44
Instructor 3	04:19	Instructor 3	04:39
Instructor 4	05:09	Instructor 4	04:27
Instructor 5	03:36	Instructor 5	03:59
Instructor 6	04:48	Instructor 6	03:58
Instructor 7	03:37	Instructor 7	05:40
Instructor 8	03:23	Instructor 8	05:25
Instructor 9	04:44	Instructor 9	03:48
Instructor 10	04:28	Instructor 10	04:21
		Instructor 11	04:28
		Instructor 12	04:54
Group Total:	42:08	Group Total:	58:83
Grand Total: 100:91 minutes			

After the standardization phase was finished, the researcher had to specify suitable date and time periods in order to conduct the interviews because of the busy teaching schedule of the instructors. Therefore, the interviews could not be performed right after the standardization sessions. Also, the interviews were done in Turkish and one-to-one with each instructor. At the beginning of the interviews, the researcher greeted instructors and thanked them for their contributions to the study. Then, he informed them that their responses would be audio-recorded and used only by the researcher for research purposes. It was also added that their responses would not be given or listened by another person. The researcher read the information on the interview document stating that the instructors could leave the interview any time when they feel uncomfortable or they did not have to answer all the questions if they did not wish to do so. After giving all the information about the interview and getting the consent of the instructor, the researcher started recording the interview. The researcher asked general questions about the training the instructors had received (See Appendix A). The idea here was to probe as many insights as possible from the instructors' experiences on both mediums of standardization. The interviews were done in a room where instructors might feel themselves comfortable and more honest.

3.7.4. Phase 3: Spoken Assessment (Official Oral Examination of the School)

In January 2019, all instructors who participated in both face-to-face and online standardization sessions were appointed in the official oral examination of the school as raters.

Table 3.8. *Assignment of Instructors in the Official Oral Examination*

Jury Number*	Assignment of Instructors**		Students scored (*N)*
Jury 1	OS8	FFS4	28
Jury 2	FFS8	OS2	32
Jury 3	FFS1	OS9	32
Jury 4	OS6	FFS12	32
Jury 5	OS7	OS1	32
Jury 6	***	***	32
Jury 7	OS11	FFS9	32
Jury 8	0S10	FFS10	32
Jury 9	FFS3	FFS2	30
Jury 10	FFS11	FFS5	28
Jury 11	OS3	OS5	30
Jury 12	FFS7	OS4	30
Jury 13	FFS6	OS12	30

* In the school context, jury refers to a particular classroom where two instructors conduct oral examinations.

**Instructors are assigned into their Juries as pairs where there is one experienced and one novice instructor from both standardization groups.

*** Instructors who did not take part in standardization sessions

Prior to the examination, the researcher assigned all instructors to the classrooms where they would rate students' oral performances. Then, he e-mailed them a document telling them where and with whom they would be rating students on the exam day.

The researcher tried to include as many combinations as possible while assigning instructors who were trained on both online and in person for the real oral examination of the school. However, as a school policy, the exam classrooms are designed in a way that there would be one experienced and one novice instructor with regard to the school's official oral exam. Therefore, some classrooms included instructors who were both trained online or face-to-face. The exam was held in 13 classrooms. 8 of the classrooms had instructors who were trained online and face-to-face, 2 of the classrooms included instructors trained face-to-face, 2 of the classrooms included instructors trained online in the standardization sessions, and 1 classroom included instructors who had not taken part in the standardization sessions (See Table 3.8.).

For the confidentiality purposes, the study does not include the names of the participants. However, the researcher gave all participants a code name which tells about their identity (See Table 3.8.). For instance, instructors trained in face-to-face standardization were coded as "FFS" and they were also given ordinal numbers so that he could understand who that number referred to. The same principle was used for the instructors who were trained on the online standardization platform. The instructors were given a code name starting with "OS", and then they were also given an ordinal number with the same purpose as in face-to-face standardization.

The instructors were first given these code names for the standardization sessions where they were required to score sample student exams. The same code names for the same instructors were used for the official examination in order to avoid any confusion about the identity of the instructor. To illustrate, if an instructor was given a code name as "OS5", he or she was given the same code name for the official speaking exam. This helped the researcher a lot while assigning instructors to their classrooms in the official oral examination of the school.

After completing the assigning of instructors, the researcher sent a document to the instructors explaining where and with whom they would be rating students on the exam day.

There were 32 students assigned to each classroom, so all instructors were expected to score 32 speaking exams. However, some students were absent on the exam day. Therefore, some instructors scored less than 32 students. Instructors in Juries, the classrooms the exam is conducted, such as Juries 2,3,4,5,7,8 managed to score 32 student exams as there were no absent students. Juries 9,11,12,13 scored 30 students, Juries 1,10 scored 28 students, and in Jury 6 instructors with no standardization training scored 32 students (See Table 3.8.)

At the end of the exam day, the instructors brought their exam packs to the testing office of the department. Then, the researcher made photocopies of the paper grading sheets that instructors used during the exam. Later, the instructor typed all grades given by the instructors on an Excel file for the data analysis

Overview of the Official Oral Exam

Spoken Assessment is the name of the oral exam of the school. It includes three parts:

- Introduction and talking about a warm-up question
- Describing a picture
- Talking about a topic

The exam is always conducted by two instructors who grade two or three students in a period of 15 minutes. In that time period, one of the instructors takes notes and communicates with the students one by one and the other instructor takes notes and time for each student. When all students finish the exam, they leave the classroom, and the instructors grade students' performances by looking at their notes and referring to the exam criteria (See Appendix C) Then, the average score is given to the test-taker as the final score. Instructors have around 5 minutes to finish grading the students until the next set of students are sent to the exam classroom by exam coordinators. The instructors score the test-takers' performance as a whole by addressing to all parts of the exam together. That is, Warm-Up and Introduction, Picture Description and Topic parts are not scored one by one. The raters score the performances by referring to all these parts as a whole. During the exam, test-takers are seated in front of raters and the board so that they can see the exam parts projected on the board.

Introduction and Talking about a Warm-up Question

In this part, raters ask each test-taker to talk about themselves briefly, and then ask an introductory question. Each test-taker is asked a different introductory question during the exam.

Describing a Picture

The aim of this stage is to describe a picture in detail and to encourage discussion about the picture to be described.

Talking about a Topic

Each test taker talks about his/her own topic within the allocated time.

All of these parts are provided to the raters as Microsoft Office PowerPoint files so that they can project them on the board during the examination. Raters are given the criteria and the marking sheet as printed documents. While test-takers are performing, they take notes and grade them by addressing to the criteria. All test-takers' performances are audio-recorded as they might be needed to be revised in case of objections or used in standardization sessions to train instructors.

3.8. Data Analysis

The data gathered for this study were the scores of sample exams graded by instructors in face-to-face standardization training and online standardization platform, post-test (official oral examination of the institute) scores of instructors trained in both mediums, and the interview data collected through semi-structured interviews with instructors trained in both mediums.

Table 3.9. *Data Analysis of the Study*

Research Questions	Method	Data Analysis
RQ1: How effective is an online standardization platform specifically designed to train teachers that will administer oral exams as raters?	Quantitative	Intraclass Correlation Coefficient (ICC), Independent Samples t-Test, Cohen's Kappa Coefficient
a. Do raters trained on an online standardization platform score oral exams consistently within their group?	Quantitative	Intraclass Correlation Coefficient (ICC)
b. Do raters trained in a face-to-face standardization training score oral exams consistently within their group?	Quantitative	Intraclass Correlation Coefficient (ICC)
c. Is there a significant difference between the scores of oral exams given by the raters trained on an online standardization platform and the ones trained in a face-to-face standardization training?	Quantitative	Independent Samples t-Test

d. Do raters trained on an online standardization platform and the ones trained in a face-to-face standardization session apply oral exam criteria consistently in an actual oral examination?	Quantitative	Cohen’s Kappa Coefficient
RQ2: What are the views of the participants about the standardization trainings they received before an official oral examination?	Qualitative	Thematic Content Analysis (NVivo v12)
a. What are the views of the participants about an online standardization platform specifically designed to train instructors that will administer oral exams as raters?	Qualitative	Thematic Content Analysis (NVivo v12)
b. What are the views of the participants about a face-to-face standardization used to train instructors that will administer oral exams as raters?	Qualitative	Thematic Content Analysis (NVivo v12)

Data Analysis of the 1st research Question (Quantitative)

In order to find answers for the 1st research question (How effective is an online standardization platform specifically designed to train teachers that will administer oral exams as raters?), several statistical analyses were implemented.

For the sub-questions a (Do raters trained on an online standardization platform score oral exams consistently within their group?) and b (Do raters trained in a face-to-face standardization training score oral exams consistently within their group?) of the 1st research question, Intraclass Correlation Coefficient (ICC) analysis was implemented (See Table 3.9.). The aim of the analysis was to provide statistical results about whether the instructors in face-to-face and online standardization groups had inter-rater reliability (agreement) with other instructors within their groups. With the ICC analysis, inter-rater reliability correlation matrix and intraclass correlation coefficient values were calculated for instructors in both groups in order to see whether there was inter-rater reliability (agreement) among instructors within their groups. The ICC was favored for this analysis as there were more than two raters (n=12) in each group. Furthermore, ICC model 3 (two-way-mixed model) was used since the subjects (10 sample exams) were assessed by each rater (*n=12)* in each standardization medium, and the raters were the only raters of the researcher's interest (Koo & Li, 2016). For the two-way-mixed model, there are two ICC definitions which are "absolute agreement" and "consistency". By addressing to the literature (Portney & Watkins, 2000; Shrout & Fleiss, 1979), absolute agreement among raters was analyzed as the inter-rater agreement for multiple scores was not rational to generalize to a larger population of raters, and measurements would not yield to any logical interpretations if there no agreement exists between repeated measurements. Portney and Watkins also emphasized that when test-retest and inter-rater agreement are to be analyzed, "absolute agreement" must be employed in order to have valid and solid outcomes.

For the sub-question c (Is there a significant difference between the scores of oral exams given by the raters trained on an online standardization platform and the ones trained in a face-to-face standardization training) of the 1st research question, the Independent Samples t-Test analysis was implemented (See Table 3.9.). The purpose of the test was to compare the mean scores of instructors in each sample exam graded by the instructors in both standardization groups. Assumptions of independence, normality, and homogeneity were confirmed (See Table 3.10).

Table 3.10. *Normality Check with Shapiro Wilk-W Analysis*

Variable	*Statistic*	*df*	*p*
Online Avg. Score	.988	10	.993*
Face-to-face Avg. Score	.991	10	.998*

* *p* >0,05

For the sub-question d (Do raters trained on an online standardization platform and the ones trained in a face-to-face standardization session apply oral exam criteria consistently in an actual oral examination?) of the 1st research question, Cohen's Kappa Coefficient (K) statistical analysis was implemented (See Table 3.9.) to see whether there was inter-rater agreement (consistency) within the pairs of instructors who rated speaking exams in the post-test (official oral examination) phase of the study. This analysis was favored since the instructors from both mediums were assigned to their classes for the official exam in pairs.

Data Analysis of the 2nd Research Question (Qualitative)

Qualitative data of the study were collected by conducting semi-structured interviews with instructors from both face-to-face and online standardization groups. The answers obtained for the interview questions were analyzed using thematic analysis on the NVivo software version 12 (See Table 3.9.). The answers were put into categories on the software under main themes. After that, the answers were analyzed and organized according to their relevancy to the 2nd research question (What are the views of the participants about an online standardization platform specifically designed to train teachers that will administer oral exams as raters?) and its sub-question (What are the views of the participants about a face-to-face standardization used to train teachers that will administer oral exams as raters?).

3.9. Trustworthiness

As this study had a mixed-method research design, the qualitative part of the study attempted gain broader insights of the participants of both training mediums. The literature suggests that qualitative studies' trustworthiness are questioned since validity and reliability concepts are not ensured the same way as quantitative ones (Krefting, 1991). Kleinsasser and Silverman (2006) emphasized that these problems can be overcome by employing some measures in order to find answers to the validity and reliability problems. These measures were described by (Schwandt, Lincoln, & Guba, 2007) as "credibility, transferability, dependability and confirmability" with regard to “internal validity, external validity/generalizability, reliability, and objectivity” respectively (Shenton, 2004).

Internal validity (credibility) of the study is concerned with the extent to which findings of the study is in accordance with the reality. According to Merriam (1998), the internal validity of a study could be ensured by the triangulation of data, peer check, member checking and the clarification of biases of the researcher. The first validation strategy was the triangulation of data, which refers to the collection of evidence from various sources to analyze and support the findings of the study (Onwuegbuzie & Leech, 2007). In this study, the researcher used interviews, documents and audio-visual materials of the participants (Lincoln & Guba, 1985) and also a different research method (quantitative data) (Denzin, & Lincoln, 2005; Lincoln & Guba, 1985) in order to shed light on the research questions and enhance the quality of the research. During the semi-structured interview phase, the researcher conducted the interviews with the participants from both mediums so that there could be overlapping responses. Although the main purpose of the study was to find out whether the online standardization platform was effective or not, the researcher also conducted interviews with participants trained in face-to-face standardization so that there could be valuable data which would support the effectiveness of the online standardization platform. Moreover, the quantitative data from the standardization sessions were utilized so that there could be overlapping data to support the effectiveness of the online platform. Another strategy used by the researcher was the “negative case analysis”. While analyzing the qualitative data, some contradictory responses which had not been expected by the researcher emerged. In order to increase the internal validity of the study, the researcher reported them in the findings of the study (Bitsch, 2005; Creswell, 2013).

As for the external validity/generalizability (transferability), the researcher provided extensive details regarding the methodology and the context of the study and employed purposive sampling where he selected samples according to their specific features such as experience as raters in the official oral examination. Bitsch (2005) and Shenton (2004) emphasized that studies could be transferred to other contexts when researchers extensive details about the study and employ purposive sampling.

To ensure the validity (dependability) of the study, the researcher employed "code-recode" and "peer examination" strategies (Bitsch, 2005). While analyzing the qualitative data, the researcher coded the same interview data in different times to decide on the final themes to be used in the research report. The researcher also consulted to one of his colleagues to do the thematic analysis on NVivo software in order to come up with the same themes so that they could be used in the results of the study. Some techniques regarding the trustworthiness could not be utilized in the study. They are mentioned in the limitation and delimitations part of the study.

3.10. Limitations and Delimitations of the Study

It is noteworthy that there are limitations in this study. To start with, all instructors who took part in the online training had to use their computer in their lessons, so they were assumed to have enough technological literacy to receive online training. However, some of these instructors would not really have enough familiarity with technology or prefer receiving training online. The researcher had not tested their technological literacy before the sampling. Therefore, the results of the study might have changed if the instructors on the online standardization platform had been chosen according to their technological literacy, preference, or the result of a pilot test. Secondly, the instructors on the online standardization platform might not have given enough effort as their colleagues in face-to-face standardization. As they were not monitored by a trainer during their training, it might have affected the results of the study. In addition, these instructors usually completed their trainings at home, so they might have been distracted or cognitively overloaded as opposed to their peers in face-to-face standardization. Lastly, the researcher knew all participants in person since they all work in the same institution. Therefore, the validity and reliability of the data provided by the participants are confined to the honesty and beliefs of the participants about the study and researcher. This condition might have also affected the outcome of the study. Thirdly, the sampling was not random and the sample size (*n=24*) was not big enough to generalize the results to the whole population of instructors who score oral examinations. Working with a larger size of samples might have changed the results drastically where the results might have been more positive or more negative regarding the aim of the study.

As a delimitation for the study, the utilization of purposive sampling might result in problems such as error of judgment since the researcher assigned the participants into particular groups based on his belief that the sampling would provide the correct data for the study. Therefore, an error in the researcher's judgment might impede the quality of the study in terms of representativeness of the selected participants and their anticipated knowledge as the sample. Another delimitation is the analyses made in the study. The statistical analyses conducted in the study focused on the inter-rater reliability (agreement) among the participants as they could provide answers for the research questions of the study. However, the researcher could have also used other analyses to provide a deeper understanding for the research questions focusing on the reliability (agreement) between and among instructors such as the factor of bias, participants' background knowledge, cultural differences among instructors, and use of grading criteria regarding "leniency and severity" (Yan, 2014). The last delimitation of the study was the measures performed to ensure the trustworthiness of the study. Although the researcher employed several measures to ensure the internal validity, generalizability, and the dependability, he did not perform measures such as "member check" in which transcriptions of interviews are provided to participants so that they could see how well and correct the transcribed documents were.

CHAPTER 4

RESULTS

This chapter analyzes the quantitative and qualitative data gathered in the study. The results of the study start with the quantitative data analysis and findings supported by various tables and figures representing the statistical results of the data. As the next part, the qualitative data analysis provides findings regarding the data collection tool used in the study. The qualitative findings of this study are supported by quotes from the participants with relevance to particular results throughout this chapter.

4.1. Quantitative Findings

This section provides the quantitative findings for the research questions below.

RQ1: How effective is an online standardization platform specifically designed to train teachers that will administer oral exams as raters?

- **a.** Do raters trained on an online standardization platform score oral exams consistently within their group?
- **b.** Do raters trained in a face-to-face standardization training score oral exams consistently within their group?
- **c.** Is there a significant difference between the scores of oral exams given by the raters trained on an online standardization platform and the ones trained in a face-to-face standardization training?
- **d.** Do raters trained on an online standardization platform and the ones trained in a face-to-face standardization session apply oral exam criteria consistently in an actual oral examination?

4.1.1. Do raters trained in an online standardization platform score oral exams consistently within their group?

The fact that the instructors were to be standardized for the official oral examination of the school required them to refresh their knowledge about the exam in general and score sample exams which were actually real student examinations of the previous academic years. Half of the instructors trained on the online standardization program had neither conducted any oral exams in the institution nor received any training regarding the official oral exam. The other half of the instructors had conducted their last oral examination at the beginning of the academic year. Hence, the researcher expected the 12 instructors to have agreement differences in grading the first three sample exams naturally.

Table 4.1. *ICC Values and their Interpretations*

Interpretation of ICC Values (Koo & Li, 2016)	
ICC Values	Interpretation
< 0,50	Poor agreement
Between 0,5 and 0,75	Moderate agreement
Between 0,75 and 0,90	Good agreement
> 0,90	Excellent agreement

According to Koo and Li (2016), in a reliability (agreement) analysis, ICC values less than .5 indicate a "poor agreement" level between or among raters, values between .5 and .75 demonstrate a "moderate agreement", values between .75 and .9 are considered as "good agreement" and values more than .9 are accepted as "excellent agreement".

Analysis of the First Three Sample Exams

Table 4.2. *Inter-Rater Agreement of the Instructors Trained Online (First Three Sample Exams)*

	OS1	OS2	OS3	OS4	OS5	OS6	OS7	OS8	OS9	OS10	OS11
OS2	,983										
OS3	,983	1									
OS4	,995	,996	,996								
OS5	,966	,997	,997	,987							
OS6	,884	,955	,955	,926	,975						
OS7	,139	,319	,319	,235	,391	,585					
OS8	,983	1	1	,996	,997	,955	,319				
OS9	,924	,838	,838	,882	,794	,639	-,250	,838			
OS10	-,032	,152	,152	,066	,228	,438	,985	,152	-,411		
OS11	,998	,993	,993	,999	,980	,912	,201	,993	,898	,031	
OS12	,892	,960	,960	,932	,979	1	,572	,960	,651	,424	,918

The inter-reliability matrix of the 12 instructors on the online standardization platform regarding the first three sample exams supported the researcher’s idea that there might be agreement differences among the instructors. As it can be seen from the matrix table, there are some instructors (OS7=experienced, OS8=experienced, OS9=novice, OS10=experienced) who had more than three poor agreement values as opposed to other instructors. OS7 and OS10 reported parallel responses to their ICC scores in semi-structured interviews as well.

> *Sometimes my grades and the ones on the system were different. At those times, I couldn’t ask anybody. Maybe we could have combined face-*to-*face meetings with you with the online training. [OS7]*
>
> *As we did everything online, I had some questions on my mind. Maybe we could have graded some sample exams before we received online feedbacks. [OS10]*

The lowest agreement measure was -.411 between instructors (OS9=novice, OS10=experienced) and the highest one was 1 among instructors (OS2=novice, OS3=novice, 0S8=experienced), and between instructors (OS6=experienced, OS12=experienced). That is, these instructors gave the same scores in the first three sample exams.

Table 4.3. *Consistency (Agreement) Values of the Instructors in the First Three Sample Exams*

Intraclass Correlation Coefficient of the First Three Sample Exams					
		95% Confidence Interval		F Test with True Value 0 (df1=2 df2=22)	
	Intraclass Correlation	Lower Bound	Upper Bound	Value	Sig
Single Measures	,750*	,382	,992	36,940	,000
Average Measures	,973*	,881	,999	36,940	,000

*Two-way-mixed model (absolute agreement)

As for the single rater measurements, The ICC (3,12) measurement was .750, with a 95% confidence interval between .382 and .992 (F (2,22) = 36.940, p<.001 suggesting that the real ICC value would be between .382 and .992; hence the agreement level is estimated to be “poor to excellent”. Although there were several agreement differences in the correlation matrix, there was a significant level of agreement among the instructors in scoring of the first three sample exams. The average ICC (3,12) measurement was .973, with a 95% confidence interval between .881 and .999 (F (2,22) = 36.940, p<.001 suggesting that the real ICC value would be between .881 and .999; hence the agreement level is estimated to be “good to excellent”.

Analysis of the First Six Sample Exams

Table 4.4. *Inter-Rater Agreement of the Instructors Trained Online (First Six Sample Exams)*

	OS1	OS2	OS3	OS4	OS5	OS6	OS7	OS8	OS9	OS10	OS11
OS2	,964										
OS3	,658	,752									
OS4	,936	,909	,795								
OS5	,763	,734	,858	,891							
OS6	,928	,975	,722	,865	,743						
OS7	,577	,478	,409	,646	,787	,588					
OS8	,960	,919	,751	,974	,905	,911	,743				
OS9	,591	,563	,766	,635	,864	,581	,611	,710			
OS10	,614	,580	,487	,601	,761	,719	,923	,742	,705		
OS11	,852	,728	,555	,911	,856	,712	,822	,926	,603	,686	
OS12	,903	,843	,419	,846	,642	,859	,702	,883	,339	,667	,863

The inter-rater reliability matrix of the first six sample exams on the online standardization platform emphasized that there had been a considerable increase in the agreement levels of instructors. By looking at the matrix, it could be inferred that the increased degree of agreement among instructors can be seen among all instructors. The lowest agreement measure was .339 between instructors OS9 (novice) and OS12 (experienced); and the highest one was .964 between novice instructors OS1 and OS2.

Table 4.5. *Consistency (Agreement) Values of the Instructors in the First Six Sample Exams*

Intraclass Correlation Coefficient of the First Six Sample Exams					
		95% Confidence Interval		F Test with True Value 0 (df1=5 df2=55)	
	Intraclass Correlation	Lower Bound	Upper Bound	Value	Sig
Single Measures	,733*	,480	,945	33,934	,000
Average Measures	,971*	,917	,995	33,934	,000

*Two-way-mixed model (absolute agreement)

As for the single rater measurements, The ICC (3,12) measurement was .733, with a 95% confidence interval between .480 and .945 (F (5,55) = 33,934, p<.001 suggesting that the real ICC value would be between .480 and .945; thus, the agreement level is estimated to be “poor to excellent”.

When all raters are taken into consideration, the average ICC (3,12) measurement for the first six sample exams on the platform was .973, with a 95% confidence interval between .917 and .995 (F (5,55) = 33,934, p<.001 suggesting that 97% of the variance among all instructors’ average scores was real and the real average ICC value is somewhere between .917 and .995, which is considered as “excellent agreement” in the worst scenario.

Analysis of the Ten Sample Exams

Table 4.6. *Inter-Rater Agreement of the Instructors Trained Online (Ten Sample Exams)*

	OS1	OS2	OS3	OS4	OS5	OS6	OS7	OS8	OS9	OS10	OS11
OS2	,962										
OS3	,815	,872									
OS4	,937	,903	,873								
OS5	,871	,856	,921	,893							
OS6	,935	,956	,878	,853	,891						
OS7	,754	,726	,732	,803	,870	,771					
OS8	,943	,910	,833	,972	,890	,857	,860				
OS9	,714	,679	,833	,752	,885	,724	,795	,794			
OS10	,807	,839	,809	,764	,889	,891	,928	,838	,808		
OS11	,887	,800	,757	,935	,892	,794	,904	,940	,770	,812	
OS12	,925	,905	,697	,907	,764	,844	,812	,933	,564	,794	,885

The inter-rater correlation matrix of the ten sample exams manifested that the level of consistency among instructors increased considerably throughout scoring the ten sample exams on the online standardization platform. The lowest measure of agreement among instructors was between instructors OS9 (novice) and OS12 (experienced) with an ICC measure of .564, and the strongest agreement among instructors was between the instructors OS4 (novice) and OS8 (experienced) with an ICC measure of .982. By looking at the matrix table, it can be understood that all ICC (3,12) measures among instructors were in the range of “good” and “excellent”.

Table 4.7. *Consistency (Agreement) Values of the Instructors in the Ten Sample Exams*

Intraclass Correlation Coefficient of the Ten Sample Exams					
		95% Confidence Interval		F Test with True Value 0 (df1=9 df2=99)	
	Intraclass Correlation	Lower Bound	Upper Bound	Value	Sig
Single Measures	,839*	,695	,947	63,496	,000
Average Measures	,984*	,965	,995	63,496	,000

*Two-way-mixed model (absolute agreement)

As for the single rater measurements, The ICC (3,12) measurement was .839, with a 95% confidence interval between .695 and .947 (F (9,99) = 63,496, p<.001 suggesting that the real ICC value would be between .695 and .947; thus, the agreement level is estimated to be "good to excellent". In accordance with the correlation matrix, the ICC table supported the idea that there was a significant level of agreement among the instructors in scoring a total of ten sample exams. The average ICC (3,12) measurement was .984, with a 95% confidence interval between .965 and .995 (F (9,99) = 63.496, p<.001 suggesting that 98% of the variance among all instructors' average scores was real and the real ICC values could land in some point between .965 and .995, which means "excellent agreement" in the worst-case scenario.

Summary of the Findings for Research Question 1a

Table 4.8. *Overview of the ICC Values of Instructors Trained on the Online Standardization Platform*

Overview of the ICC of the Online Standardization Group			
Exams	ICC*	95% Confidence Interval	
		Lower Bound	Upper Bound
First Three Exams	Single: ,750	,382	,992
	Average: ,973	,881	,999
First Six Exams	Single: ,733	,480	,945
	Average: ,971	,917	,995
Ten Sample Exams	Single: ,839	,695	,947
	Average: ,984	,965	,995

*Two-way-mixed model (absolute agreement)

The findings for the sub-question (a) of the 1st research question (Do raters trained on an online standardization platform score oral exams consistently within their group?) suggested that the consistency among instructors on online standardization platform showed increased average ICC (3,12) measures of .973, .971, and .984 in scoring the first three, first six, and all ten sample exams respectively. In addition, the individual ICC measures also increased with the ICC (3,12) measures of .750, .733 and .839 in scoring the first three, first six, and all ten sample exams respectively. The online standardization group demonstrated low levels of correlation among each other in the first three sample exams. However, as they progressed, the agreement levels among them showed a considerable increase.

As seen from the Table 4.8., the average ICC (3,12) measures of the first three, first six, and the total number of the sample exams were .973, .971, and .984 respectively, which indicated "excellent agreement" among instructors. Also, the 95% confidence interval values of the single measures demonstrated a considerable increase, which meant that there was 95% chance that true ICC measure would fall on any value between lower bound and upper bound. While in the first three exams, these true values fall in between .382 and .992, in the next exams these values increased considerably .480 and .945 in the first six sample exams, and .695 and .995 in all of the ten sample exams combined.

4.1.2. Do raters trained in a face-to-face standardization training score oral exams consistently within their group?

Considering the fact that the instructors were to be standardized for the official oral examination of the school, they were supposed to refresh their knowledge about the exam in general and score sample exams which were actually real student examinations of the previous academic years. Having known that half of these instructors had not conducted any oral exams in the institution, and the other half had conducted their last examination at the beginning of the academic year, the researcher expected them to have agreement differences in grading the first three sample exams naturally.

According to Koo and Li (2016), in a reliability (agreement) analysis, ICC values less than .5 indicate a "poor agreement" level between or among raters, values between .5 and .75 demonstrate a "moderate agreement", values between .75 and .9 are considered as "good agreement" and values more than .9 are accepted as "excellent agreement" (See Table 4.1).

Analysis of the First Three Sample Exams

Table 4.9. *Inter-Rater Agreement of the Instructors Trained Face-to-face (First Three Sample Exams)*

	FFS 1	FFS 2	FFS 3	FFS 4	FFS 5	FFS 6	FFS 7	FFS 8	FFS 9	FFS 10	FFS 11
FFS2	,500										
FFS3	,768	,939									
FFS4	1	,500	,768								
FFS5	,906	,820	,967	,906							
FFS6	,956	,731	,922	,956	,990						
FFS7	,986	,635	,863	,986	,963	,991					
FFS8	,990	,615	,849	,990	,956	,988	1				
FFS9	,500	1	,939	,500	,820	,731	,635	,615			
FFS10	,986	,635	,863	,986	,963	,991	1	1	,635		
FFS11	,799	,920	,999	,799	,979	,940	,887	,874	,920	,887	
FFS12	,928	,787	,952	,928	,998	,996	,976	,970	,787	,976	,966

The inter-reliability matrix of the 12 instructors in face-to-face standardization supported the researcher's idea in the same way. As it can be seen from the matrix table, there are some instructors (FFS2=novice, FFS9=experienced) who had multiple poor ICC (3,12) scores in relation to other instructors.

Table 4.10. *Consistency (Agreement) Values of the Instructors in the First Three Sample Exams*

Intraclass Correlation Coefficient of the First Three Sample Exams					
		95% Confidence Interval		F Test with True Value 0 (df1=2 df2=22)	
	Intraclass Correlation	Lower Bound	Upper Bound	Value	Sig
Single Measures	,807*	,481	,994	43,051	,000
Average Measures	,980*	,917	1,000	43,051	,000

*Two-way-mixed model (absolute agreement)

When the single rater measurements are taken into account, the ICC (3,12) average measure of the first three sample exams was .807, with a 95% confidence interval between .481 and .994 (F (2,22) = 43.051, p<.001). That is, the real ICC values for single raters would land somewhere between .481 and .994, which means "poor to excellent agreement". Overall, it can be seen that there was a significant level of agreement among instructors. The ICC (3,12) average measure of the first three sample exams was .980, with a 95% confidence interval between .917 and 1 (F (2,22) = 43.051, p<.001). It can be inferred from the average measure that 98% of the variance of the overall averaged scores of the instructors is real and the real ICC values would land somewhere between .917 and 1, which means an "excellent agreement" in the worst-case scenario. Throughout the grading of the first three sample exams, the highest degree of reliability was seen between instructors (FFS1=experienced, FFS4=novice), (FFS2=novice, FFS9=experienced) and instructors (FFS7=experienced, FFS8=experienced, FFS10=novice) with an ICC of 1, which means that these instructors gave the same scores for all the first three sample exams. Conversely, the lowest level of reliability was seen between the instructors (FFS1=experienced, FFS2=novice), (FFS1=experienced, FFS9=experienced), and (FFS4=novice, FFS9=experienced) with an ICC measure of .500.

Analysis of the First Six Sample Exams

Table 4.11. *Inter-Rater Agreement of the Instructors Trained Face-to-face (First Six Sample Exams)*

	FFS 1	FFS 2	FFS 3	FFS 4	FFS 5	FFS 6	FFS 7	FFS 8	FFS 9	FFS 10	FFS 11
FFS2	,875										
FFS3	,928	,965									
FFS4	,948	,833	,873								
FFS5	,816	,878	,855	,922							
FFS6	,978	,933	,977	,953	,890						
FFS7	,838	,816	,858	,944	,970	,905					
FFS8	,851	,833	,856	,957	,967	,909	,978				
FFS9	,803	,935	,964	,734	,775	,890	,771	,773			
FFS10	,896	,863	,927	,946	,938	,955	,982	,954	,851		
FFS11	,727	,893	,894	,787	,922	,852	,908	,885	,909	,916	
FFS12	,891	,853	,848	,948	,945	,913	,928	,898	,706	,920	,811

The inter-rater agreement matrix of the first six sample exams showed that there had been a considerable increase in the agreement levels of instructors. By looking at the matrix, it could be noted that the degree of agreement among instructors can be seen among all instructors. Analyzing the matrix from instructor to instructor, it is seen that FFS9 consistently had lower correlation scores as opposed to other instructors. The minimum correlation measure was .706 between instructors (FFS9=experienced, FFS12=novice) and the highest one was .982 between instructors (FFS7=experienced, FFS10=novice).

Table 4.12. *Consistency (Agreement) Values of the Instructors in the First Six Sample Exams*

Intraclass Correlation Coefficient of the First Six Sample Exams					
		95% Confidence Interval		F Test with True Value 0 (df1=5 df2=55)	
	Intraclass Correlation	Lower Bound	Upper Bound	Value	Sig
Single Measures	,863*	,690	,975	76,780	,000
Average Measures	,987*	,964	,998	76,780	,000

*Two-way-mixed model (absolute agreement)

When the single rater measurements are considered, the ICC (3,12) single measures of the first six sample exams was .863, with a 95% confidence interval between .690 and .975 (F (5,55) = 76,780, p<.001). That is, the real ICC values for single raters would land somewhere between .690 and .975, which means "good to excellent agreement".

Overall, the results of the ICC analysis of the first six sample exams showed that the agreement among instructors throughout grading the six sample exams was significant. The level of agreement among instructors can be summarized with average measures of ICC (3,12) of .987, with a 95% confidence interval between .964 and .998 (F (5,55) =76.780, p<.001, which means that the real average ICC measures were within the "excellent agreement" range.

Analysis of the Ten Sample Exams

Table 4.13. *Inter-Rater Agreement of the Instructors Trained Face-to-face (Ten Sample Exams)*

	FFS 1	FFS 2	FFS 3	FFS 4	FFS 5	FFS 6	FFS 7	FFS 8	FFS 9	FFS 10	FFS 11
FFS2	,859										
FFS3	,865	,942									
FFS4	,915	,916	,943								
FFS5	,855	,931	,909	,960							
FFS6	,947	,902	,959	,958	,918						
FFS7	,870	,870	,867	,947	,973	,911					
FFS8	,884	,898	,889	,966	,979	,922	,985				
FFS9	,903	,929	,890	,847	,855	,888	,843	,860			
FFS10	,920	,858	,918	,941	,929	,979	,947	,938	,868		
FFS11	,898	,842	,864	,849	,871	,909	,858	,865	,924	,929	
FFS12	,882	,887	,881	,954	,962	,926	,954	,938	,795	,927	,815

The inter-rater correlation matrix of the ten sample exams emphasized that the consistency among instructors increased remarkably throughout scoring the ten sample exams in the standardization training. The weakest agreement among instructors was between instructors FFS9 (experienced) and FFS12 (novice) with an ICC measure of .795, and the strongest agreement among instructors was between the instructors FFS7 (experienced) and FFS8 (experienced) with an ICC measure of .985. By looking at the matrix table, it can be understood that all ICC (3,12) measures among instructors were in the range of “good” and “excellent”.

Table 4.14. *Consistency (Agreement) Values of the Instructors in the Ten Sample Exams*

Intraclass Correlation Coefficient of the Ten Sample Exams					
		95% Confidence Interval		F Test with True Value 0 (df1=9 df2=99)	
	Intraclass Correlation	Lower Bound	Upper Bound	Value	Sig
Single Measures	,891*	,782	,965	98,784	,000
Average Measures	,990*	,977	,997	98,784	,000

*Two-way-mixed model (absolute agreement)

As for the single rater measurements, the ICC (3,12) single measures of the ten sample exams was .891, with a 95% confidence interval between .782 and .965 (F (9,99) = 98,784, p<.001). That is, the real ICC values for single raters would land somewhere between .782 and .965, which means "good to excellent agreement". In accordance with the correlation matrix, the intraclass correlation coefficient table supported the idea that there was a significant level of agreement among the instructors in scoring a total of ten sample exams. The average ICC (3,12) measurement was .990, with a 95% confidence interval between .977 and .997 (F (9,99) = 98.784, p<.001 suggesting that the real average ICC value is somewhere between .977 and .997, which means "excellent agreement" in the worst-case scenario.

Summary of the Findings for Research Question 1b

Table 4.15. *Overview of ICC Values of Instructors in Face-to-face Standardization*

Overview of the ICC of the Face-to-face Standardization Group			
Exams	ICC*	95% Confidence Interval	
		Lower Bound	Upper Bound
First Three Exams	Single: ,807	,481	,994
	Average: ,980	,917	1,000
First Six Exams	Single: ,863	,690	,975
	Average: ,987	,964	,998
Ten Sample Exams	Single: ,891	,782	,965
	Average: ,990	,977	,997

*Two-way-mixed model (absolute agreement)

The findings for the 1st research question (Do raters trained in a face-to-face standardization training score oral exams consistently within their group?) suggested that the agreement levels among instructors had gradually increased with the average ICC (3,12) measures of .980, .987, and .990 in scoring the first three, first six, and all ten sample exams respectively. In addition, the individual ICC measures also increased with the ICC (3,12) measures of .807, .863 and .891 in scoring the first three, first six, and all ten sample exams respectively. Starting from the correlation matrix of grading the first three sample exams to the first six and total of ten sample exams, instructor FFS9 (experienced) demonstrated the weakest consistency with other instructors. However, the instructors' correlation measures with other instructors also increased throughout grading the total of ten sample exams. As seen from the table 4.15, the average ICC (3,12) measures of the first three, first six, and the total number of the sample exams were .980, .987, and .990 respectively, which indicated excellent reliability among instructors. Also, the 95% confidence interval values of the single measures demonstrated a considerable increase, which meant that there was 95% chance that true ICC measure would fall on any value between lower bound and upper bound. While in the first three exams, these true values fall in between .481 and .994, in the next exams these values increased considerably .690 and .975 in the first six sample exams, and .782 and .965 in all of the ten sample exams combined.

Comparison of the ICC measures of Online Standardization and Face-to-face Standardization Groups

Table 4.16. *Comparison of ICC Values of Online and Face-to-face Standardization Groups*

Comparison of the ICC of Online and Face-to-face Standardization Groups			
Groups	ICC*	95% Confidence Interval	
		Lower Bound	Upper Bound
Online	Single: ,839	,695	,947
	Average: ,984	,965	,995
Face-to-face	Single: ,891	,782	,965
	Average: ,990	,977	,997

*Two-way-mixed model (absolute agreement)

Overall, both standardization groups showed an excellent level of agreement within their groups in scoring ten sample exams. Therefore, it could be accepted that the instructors on the online standardization platform and the ones in face-to-face standardization group scored sample speaking exams consistently within their groups. Both groups' average measures were above .900, which fell in to the range of “excellent” agreement. However, the comparison of the two groups demonstrated that the instructors trained in face-to-face standardization had higher agreement values than the ones on the online standardization platform.

4.1.3. Is there a significant difference between the scores of oral exams given by the raters trained on an online standardization platform and the ones trained in a face-to-face standardization training?

The third research question attempted to discover whether there was any significant difference between face-to-face standardization and online standardization groups in terms of scoring ten sample oral exams.

Table 4.17. *Independent Samples t-Test Analysis of Online and Face-to-face Standardization Groups*

Exam	Group	M	SD	*n*	*t*	*p*
Exam1	Face-to-face	53,50	7,994	12	-1,347	,192*
	Online	58,75	10,881	12		
Exam2	Face-to-face	41,25	6,355	12	-1,645	,114*
	Online	45,08	4,981	12		
Exam3	Face-to-face	73,42	8,969	12	-827	,417*
	Online	77,25	13,315	12		
Exam4	Face-to-face	59,92	7,141	12	-1,237	,229*
	Online	65,58	14,171	12		
Exam5	Face-to-face	71,08	6,501	12	-,803	,431*
	Online	73,92	10,352	12		
Exam6	Face-to-face	93,92	5,869	12	-1,909	,069*
	Online	97,67	3,447	12		
Exam7	Face-to-face	48,75	9,156	12	,294	,771*
	Online	47,67	8,886	12		
Exam8	Face-to-face	34,92	7,090	12	-2,311	,031
	Online	42,33	8,563	12		
Exam9	Face-to-face	58,75	8,433	12	,690	,497*
	Online	61,00	7,508	12		
Exam10	Face-to-face	18,25	6,384	12	-1,021	,318*
	Online	20,58	4,680	12		

* p > .05.

As seen from Table 4.17., for each of the exams (*n=10*), the mean scores of the instructors in both online and face-to-face standardization groups were compared with independent samples t-test. The results of the test demonstrated that there were no statistical differences in the mean scores of nine sample exams between two groups. However, there was a significant difference between online and face-to-face standardization groups in the exam 8, *t* (22) = -2.31, p = .31. From these results, it could be concluded that the two standardization groups provided similar scores for nine of the exams out of ten.

4.1.4. Do raters trained on an online standardization platform and the ones trained in a face-to-face standardization training apply oral exam criteria consistently in an actual oral examination?

For the 4th research question, the researcher decided to use Cohen's Kappa statistic as it is robust in testing the inter-reliability of two raters. The Kappa values range between -1 and +1.

Table 4.18. *Kappa Values and their Interpretations*

Interpretation of Kappa Values (McHugh, 2012)	
Kappa Values	Interpretation
≤ 0	No agreement
Between 0,01 and 0,19	Slight agreement
Between 0,20 and 0,39	Fair agreement
Between 0,40 and 0,59	Moderate agreement
Between 0,60 and 0,79	Substantial agreement
Between 0,80 and 0,90	Excellent agreement
> 0,90	Perfect agreement

According to McHugh (2012), Kappa values equal to or below "0" shows that there is "no agreement", .01 to .19 no to "slight agreement", .20 to .39 "fair agreement", .40 to .59 "moderate agreement", .60 to .79 "substantial agreement", .80 to .90 "excellent agreement", and above .90 "perfect agreement".

Table 4.19. *Kappa Statistics of the Pairs in the Official Oral Examination*

Pairs	K	Asymptotic SE	Approx. T	Approx. Sig.
1. OS8-FFS4	**,603** *n* = 28	,098	9,911	,000
2. OS2-FFS8	**,638** *n* = 32	,091	9,653	,000
3. OS9-FFS1	***,617*** *n* = 32	,091	10,669	,000
4. OS6-FFS12	*,183* *n* = 32	,086	3,029	,002
5. OS1-OS7	***,714*** *n* = 32	,088	11,195	,000
6. *	*,077* *n* = 32	,079	1,223	,221
7. OS11-FFS9	***,646*** *n* = 32	,091	10,185	,000
8. OS10-FFS10	***,617*** *n* = 32	,091	10,535	,000
9. FFS2-FFS3	*,301* *n* = 30	0,92	5,326	,000
10. FFS5-FFS11	*,592* *n* = 28	,103	8,364	,000
11. OS3-OS5	*,205* *n* = 28	,093	3,319	,001
12. OS4-FFS7	***,706*** *n* = 30	,089	11,717	,000
13. OS12-FFS6	***,626*** *n* = 30	,095	10,069	,000

* Pair of instructors who did not participate in any of the standardization mediums

As seen from Table 4.19., eight pairs (Pair 1, 2, 3, 5, 7, 8, 12, 13) of instructors who rated students in the official oral examination represented "substantial agreement" with K values ranging from .603 to .714. The agreement between instructors in Pair 10 demonstrated a "moderate agreement" with a K value of .592. Pairs 9 and 11 showed a "fair agreement" in their groups with K values of .301 and .205 respectively.

Pairs 4 and 6 manifested “very little agreement” in their groups with K values of .183 and .077 respectively. The highest K value belonged to Pair 5 with a value of .714 while the lowest K value was obtained from Pair 6 with a value of .077. Pair 6 consisted of raters who did not receive any standardization trainings.

From these findings, it could be noted that eight pairs of instructors out of thirteen pairs agreed with their partner substantially in the official oral examination. The instructors in one pair agreed with each other moderately. Two pairs of instructors agreed with their partners fairly. Lastly, two pairs of instructors had very little agreement (consistency) with their partners while scoring the student exams in the official oral examination.

Table 4.20. *Agreement Percentages Among Raters in the Official Oral Examination*

Pairs of Raters	N (Ratees)	Exact (0)**	Near (1-7)**	Discrepant (8-13)**	Contradictory (20 and more)**
OS8-FFS4	28	64,29%	25,00%	10,71%	0
OS2-FFS8	32	62,50%	37,50%	0	0
OS9-FFS1	32	65,63%	34,38%	0	0
OS6-FFS12	32	28,13%	59,38%	12,50%	0
OS1-OS7	32	75%	25,00%	0	0
*	32	18,75%	59,38%	12,50%	9,38%
OS11-FFS9	32	68,75%	28,13%	0	3,13%
OS10-FFS10	32	50%	43,75%	6,25%	0
FFS2-FFS3	30	36,67%	60%	3,33%	0
FFS5-FFS11	28	64,29%	35,71%	0,00%	0
OS3-OS5	28	28,57%	67,86%	3,57%	0
OS4-FFS7	30	73,33%	26,67%	0	0
OS12-FFS6	30	60%	30%	10%	0
Overall	398	53,53%	40,98%	4,53%	0,96%

* Pair of instructors who did not participate in any of the standardization mediums

** Score differences between instructors (out of 100)

Overall, Table 4.20. shows that instructors obtained a relatively high level of agreement among each other (Exact agreement = 54%, near agreement = %41). Also, the rest of the scores given to the ratees fall into the discrepant (5%) and contradictory (%1) agreement. According to the school policy, score differences ranging from 7 to 15 are accepted in the oral examinations. That is, there might be instructors whose scores differ between 7 and 15 points, and they do not have to reconsider their scores because the band where the students' proficiency levels are described do not change when there is a difference of scores between 7 and 15 points (See Appendix C). When there is a score difference of 20 points or more, the band where students' proficiency levels are described changes, so the raters have to reconsider their scores by going over their notes and discussing the score of the student with their partners. Finally, a particular student gets the average score given by the two instructors as their final score. Considering the institution's grading policy, it could be said that there was an impressive level of agreement among all instructors with around %95.

4.2. Qualitative Findings

This section provides the quantitative findings for the research questions below.

RQ2: What are the views of the participants about the standardization trainings they received before an official oral examination?

- **a.** What are the views of the participants about an online standardization platform specifically designed to train instructors that will administer oral exams as raters?
- **b.** What are the views of the participants about a face-to-face standardization used to train instructors that will administer oral exams as raters?

In this section, qualitative data findings from the semi-structured interviews are presented. The results obtained from the qualitative data analysis provide insights, opinions, and experiences of the participants in a thematic way. This section is divided into two sub-sections (Positive aspects and Concerns & Suggestions) with regard to the interview questions. Under each sub-section themes and patterns extracted from NVivo software are introduced as bulleted lists supported with relevant quotes from the participants.

Table 4.21. *Themes Extracted from the Content Analysis of Semi-structured Interviews*

	Themes Extracted from the Thematic Content Analysis	
Positive Aspects	• Training Content (Both mediums) - General information about the exam, exam procedures and rules - Selection of the sample exams - Videos used for sample speaking exams • Implementation of the Training (Both mediums) - Feedbacks - Duration of the training • Design of the Online Platform (Online) - Instructional design (Layout, step-by-step instruction, navigation, colors and coding) - Practicality & Flexibility	
Concerns & Suggestions	Online Standardization	Face-to-face Standardization
	- Need for a discussion environment - Need for combining online training with face-to-face training	- Implementing the training in longer periods - Having discussion and feedback after each sample exam

In relation to the interview questions, the first subsection is comprised of findings regarding the positive aspects of the standardization training mediums. The second subsection deals with the concerns, weaknesses, and suggestions for the improvement of the training mediums.

4.2.1. Positive Aspects about the Training Mediums

Several positive aspects concerning online standardization platform and face-to-face standardization training were posed by the instructors. While instructors trained in person mentioned that the videos in used for sample speaking exams and feedbacks given during the training were really beneficial for them, the instructors trained online mostly addressed to the design of the platform, practicality of the platform and the videos used to cover information about the exam as the strongest features of the platform. According to these main themes, this section covers several sub-themes related to online standardization, face-to-face standardization, and both as bulleted lists.

- ***Training Content***

Instructors in both standardization mediums provided positive feedbacks related to the content of the standardization. Several instructors stated that the training content on exam procedures, exam rules and general information about the exam were quite practical and informative. Some selected quotations regarding this theme are as follows:

> *It was really good to receive general information about the exam before we started to grade sample exams. [FFS10]*
>
> *I had a chance to learn about the exam procedures, how to grade student exams, what procedures to follow before-during-after the exam. I found this quite useful. Everything was clear and easy to follow. [OS1]*
>
> *I had a lot of questions on my mind before this standardization, but I was able to get the information I needed from the videos. [OS4]*
>
> *I think there wasn't any problem with the content. Everything was clear and easy to follow. There weren't extra details and sets of information. [OS5]*
>
> *Speaking frankly, I quite enjoyed the content. I know everything about what to do from the beginning of the exam until the end of it because this standardization gave me all I needed. [OS8]*

The only difference was that instructors in face-to-face standardization received the content from PowerPoint slides read by the researcher while the ones on online standardization platform obtained the training content through videos. However, none of the instructors, regardless of their standardization medium, mentioned negative feedbacks regarding the training content. One of the most common themes regarding the content was the selection of the sample exams. Some quotes from the instructors regarding the selection of sample exams were presented below.

> *I found the sample exams quite quality. We had a chance to see extreme examples. In fact, in these exams, we could see the things we could expect from our students. We could also see what grades were given to these students. This is a quite useful thing that will make our jobs easier. [FFS1]*
>
> *I liked the idea that all sample exams were chosen from different English proficiency levels. If they had been all the same, I would have been very bored and it wouldn't have had any positive effect on us. So, that the selection of the students with different mistakes and problems was really good. [FFS4]*
>
> *We first saw sample exams, then we understood how to grade these exams. [FFS5]*

Most of the instructors trained on online standardization platform and in face-to-face standardization sessions stated that videos were one of the best features of the standardization training. Firstly, the majority of instructors from both mediums praised videos used for the sample speaking exams since they knew which exam part (Introduction & Warm-up, Describing a Picture, Talking about a Topic) the student was talking about, what questions the students were answering and which pictures the students were describing by just looking at the video. Some quotations related to videos are presented below.

> *I think the way the videos were prepared and inserting pictures while students were describing pictures were really good. [OS1]*
>
> *Videos show us the parts of the exam such as warm-up. Then, we can see the picture like we are in a real exam. So, giving that reality to us is really nice. In previous years, we just listened to sample exam audios. That was much more difficult for us. Here, it was as if the students were sitting in front of us, so we gave more appropriate grades. [FFS6]*
>
> *Presenting the exams in a video was really good. It was great to see what the students were talking about and which picture they were describing while they were speaking. [FFS8]*
>
> *That the questions students were answering, the pictures, and the topics could be seen in the video was really effective. [FFS10]*

Another positive aspect mentioned by the instructors trained online regarding the videos was that they could get trained on the general information about the exam, exam procedures, and exam rules effectively.

> *For instance, the videos were really good. The ones that explain things about the exam. They weren't too long and they didn't extra information. So, there wasn't any redundant information. I can say that all of them were to the point. [OS4]*
>
> *Videos were really well-prepared. [OS5]*
>
> *The videos didn't have verbal explanations. There were just texts. It was easier for me to read the information and learn them. It was nice in that way. [OS7]*

- ***Implementation of the Training***

Regarding the implementation of the standardization trainings, both groups reported positive feelings with regard to the feedbacks during the trainings and the duration of the trainings. Instructors in both standardization mediums asserted that receiving feedbacks after sample exams was beneficial for them. Most of them stated that understanding why a particular student should get a particular grade was vital for them as it gave them the confidence needed to grade student exams reliably. Also, feedbacks were the most common theme in the analysis of the interviews by being mentioned twenty times. Some quotations regarding the feedbacks are as follows.

> *In the feedback I received, it was clearly shown that this student should get this grade because of these reasons, so I need to focus on these key points. These feedbacks were quite informative. [OS1]*
>
> *Explaining everything one by one in the feedbacks was helpful. I could understand it better with the examples better. [OS9]*
>
> *Seeing the feedback pages after grading students was useful because I understood what I should focus on while grading the students. [OS10]*
>
> *Being able to receive feedbacks after each sample exam was nice. Seeing what to focus on and how we should grade students was really nice. [OS11]*
>
> *Thinking over the exam as a group and receiving feedback in relation to that helped me. If there weren't something like this, I would be confused. [FFS2]*

> *The feedbacks you gave after the 3 sample graded exams guided us well. So, receiving feedback after each exam would be really nice, but it could cause some problems regarding the duration. [FFS3]*
>
> *I wish we could have discussed the exam more, but we couldn't. [FFS4]*
>
> *Feedbacks were one of the most important things for me. They were really effective. [FFS7]*
>
> *Having an environment where we could discuss was really nice. After all, we need to hear what other instructors are thinking in order to be standardized. Therefore, speaking about our thoughts about the grades is more important. We could have discussed the exams that we did not talk about in the sessions. [FFS9]*

As for the duration of the trainings, instructors trained on the online standardization platform indicated that they did not have any problems regarding the time they spent on the platform as they could direct their own training themselves. The instructors trained face-to-face generally expressed that having the standardization in two sessions was quite helpful as they would be tired and bored if it had been conducted in one training session.

> *I think having the standardization in two sessions is nice because it increases the possibility of being objective while grading. If the standardization were just one session, we could start comparing students with each other, and we would lose our motivation. [FFS2]*
>
> *It was nice to have the standardization in two separate days regarding our motivation and ability to make judgments of student exams. [FFS3]*
>
> *We had two sessions. That's the way it should be. Otherwise, we could miss vital points during the standardization if it were too long. [FFS6]*

Here, it can be seen that the justification of dividing the standardization into two sessions helped the researcher avoid negative feedbacks. As mentioned in the procedure section of the methodology chapter, the researcher had to divide face-to-face standardization into two because of the time limitations. Thus, it can be inferred that instructors' positive responses about having the standardization in two sessions shows that having the standardization in just one session would be quite long and ineffective. One of the instructors reported:

> *It was really rational to have the standardization in two sessions because we need to focus on the training. Had it been longer, it could have been problematic for the instructors. [FFS9]*

In addition to face-to-face standardization, the instructors trained online expressed that they did not have problems concerning the duration of the standardization training.

> *I think it was neither long nor short in terms of duration. It was OK. [OS1]*
>
> *It is really soothing to have a platform where we can get training whenever we want. It is a big advantage to eliminate problems regarding the duration. [OS2]*

- ***Design of the Online Standardization Platform***

The sub-themes regarding the design of the online platform was the instructional design of the platform, which deals with the elimination of extraneous materials and cognitive load, layout, step-by-step instruction, navigation, colors and coding mechanism of the platform. All instructors trained on the platform mentioned at least one positive aspect regarding the design of the platform. The first common theme regarding the design was the layout of the platform. Some instructors expressed their positive opinions about the layout. The layout is also concerned with the instructional design principles which help decrease distractive materials and cognitive overload.

> *Having the criteria on the left and the exam video on the right was practical. [OS2]*
>
> *It was good to have the criteria and the exam video on the same page. [OS6]*
>
> *Seeing the criteria next to the exam video made me comfortable. [OS8]*
>
> *The pages of the platform were full-pages and the colors were simple and beautiful. [OS10]*
>
> *There wasn't too much information on the platform, so I could get the information I needed easily. [OS12]*

Several instructors stated that navigating through the platform was easy and it was not difficult to find the information they needed since the training was given step-by-step. Therefore, it was pointed out that using the platform was easy for the instructors.

> *Giving the training step-by-step was nice. We could navigate to the previous page before submitting our grades. That was a good feature. [OS3]*
>
> *Everything is prepared step-by-step. We know what to find and where to find it. [OS5]*
>
> *Giving the training in three sections was helpful. [OS9]*
>
> *There were three sections on the platform. I liked it when we clicked on a menu, the information appeared below that menu… Giving the content of the exam and dividing the sections into three were good for me. [OS10]*

The coding mechanism working behind the grading system also received positive feedbacks. Several instructors mentioned that automatic conversion of their grades in to a grade out of hundred was really useful because they did not have to make mathematical calculations on their own like in face-to-face standardization and the real oral examinations of the school.

> *Being able to see the total grade of the student while grading was important for me. [OS1]*
>
> *The conversion of the grade we gave below into a grade out of hundred directly, and not having to deal with calculations were very good and practical. [OS2]*
>
> *I liked the grading part best. Converting the grades automatically when we give the grades according to the criteria…It is a really useful application. [OS5]*
>
> *Right after grading, it was really good to see what grade the student got out of hundred. [OS10]*

- ***Practicality and Flexibility of the Platform***

Several positive feedbacks regarding the practical use of the platform were expressed by the instructors. The main themes were being flexible and time-saving. Several instructors stated that they had no difficulty in completing the standardization training as they could decide on their own pacing and take breaks any time they wanted. Moreover, some other instructors added that they saved time as they could log on to the platform whenever they wanted.

> *It is nice to have a platform that we can reach any time we want. It is a great advantage to eliminate the time and place limitations. [OS2]*

> *Being able to reach the materials immediately helped me understand some parts. Instead of asking someone else, having the information in my hands helped me to get rid of problems. [OS4]*
>
> *The thing I liked most was that I could grade student exams whenever I wanted unlike a face-to-face standardization session. [OS5]*
>
> *Being able to take a break anytime we want is nice. [OS6]*
>
> *I think it's great. Transferring such a thing to an online platform is very time-saving and logical… Converting standardization into an online platform was a very time-saving and practical thing. [OS7]*

Moreover, another positive aspect regarding the practicality of the platform was the ability to stop the videos or navigate through the videos whenever the instructors needed.

> *I think grading students online was more effective because I could stop the videos, go back or go forward on the videos whenever I wanted. I listened to some part a couple of times. I can say that it was more practical for me. [OS3]*

Some instructors stated that they completed the standardization at home showing the flexibility of time and place of the platform.

> *Having the standardization at an online platform was nice. I spent some time at home doing it. It first bothered me, but I didn't have any confusion regarding the exam after completion. [OS7]*
>
> *That it was online was nice for me because I did it at home. I think it is a good thing that I didn't have be at school for this. [OS12]*

4.2.2. Concerns and Suggestions

This section helped the researcher to gain more insights about each of the standardization mediums because there were some concerns and suggestions regarding each type of standardization that were not anticipated by the researcher.

Concerns and Suggestions for the Online Standardization Platform

The most common themes addressing to this section were creating a discussion forum on the online standardization platform and combining online and face-to-face standardization training.

The first concern and suggestion by some instructors was that there could be a discussion forum on the platform where instructors could share their opinions about particular exams, answer questions of other instructors, or ask questions to others. This way, they could understand the feedbacks provided after sample exams in detail, or they could object to particular student grades explaining their reasons why those students should get a different grade. That the notion of being able to exchange ideas among instructors was seen as an important aspect to be improved by the instructors trained on the online standardization platform.

> *I want to send a message to the administrator when there is a grade I don't agree on. As the content of the sample exams are the same for everybody, I think there could be a discussion forum on the platform. [OS2]*
>
> *Sometimes my grades and the ones on the system were different. At those times, I couldn't ask anybody. [OS7]*
>
> *As we did everything online, I had some questions on my mind. [OS10]*
>
> *It is very important to have discussion. After all, we need to know what other instructors think in order to be standardized. [FFS9]*

Another suggestion by several instructors was that receiving feedbacks online was not enough; therefore, they could have received face-to-face feedbacks before they move on to grading other sample speaking exams. That is, they could start training face-to-face and get comfortable about the exam, then they could continue grading sample exams on the online platform. They also added that this way, their trainings would be more effective as they could get the necessary information from a coordinator. As a result, the theme "combining face-to-face and online standardization" emerged as a common suggestion by the instructors.

> *Receiving the feedbacks from an instructor verbally would be nice. [OS1]*
>
> *Maybe we could have combined face-to-face meetings with you with the online training… As I mentioned before, we could meet you and then continue on the online platform. [OS7]*
>
> *Maybe we could have graded some sample exams before we received online feedbacks. [OS10]*

Concerns and Suggestions for Face-to-face Standardization Platform

Some instructors from face-to-face standardization training expressed that the standardization training could have been conducted in a longer period of time so that they could grasp the information better and be less tired.

> *We could have done this training in a longer period of time. [FFS10]*

Another suggestion from the instructors trained in face-to-face standardization was that having feedbacks after each sample exam. As mentioned in the methodology chapter, the instructors received feedback after the 3 sample graded exams, but they did not receive any verbal feedback while scoring the 10 sample exams because of the time limitations. They received written feedbacks after the grading ended.

> *The feedbacks you gave after the 3 sample graded exams guided us well. So, receiving feedback after each exam would be really nice. [FFS3]*
>
> *I wish we could have discussed the exam more, but we couldn't. [FFS4]*
>
> *We could have discussed the exams that we did not talk about in the sessions. [FFS9]*

4.3. Summary of the Results

In both online and face-to-face standardization trainings, the instructors demonstrated an excellent level of agreement within their groups. The average ICC (3,12) measurement of online standardization training was .984, with a 95% confidence interval between .965 and .995 (F (9,99) = 63.496, p<.001 while the average ICC (3,12) measurement of face-to-face training was .990, with a 95% confidence interval between .977 and .997 (F (9,99) = 98.784, p<.001.

Additionally, the two standardization groups did not have significant differences in terms of scores given for the ten sample exams they graded in their standardization trainings. The mean scores of the instructors in both standardization groups were compared with independent samples t-test. The results showed that there were no statistical differences in the mean scores of nine sample exams between two groups.

However, there was a significant difference between the two standardization groups in the one of the exams, *t* (22) = -2.31, p = .31.

Furthermore, in the official oral examination, eight pairs of the instructors (9 from online standardization and 7 from face-to-face standardization) demonstrated “substantial agreement” with K values ranging from .603 to .714. The instructors in one pair agreed with one another moderately. Two pairs of instructors agreed with their partners fairly. Lastly, two pairs of instructors (one pair of instructors who did not have any standardization training, one instructor from online training, and one from face-to-face training) had very little agreement (consistency) with their partners while scoring the student exams in the official oral examination.

Finally, semi-structured interviews conducted with instructors (face-to-face *n*=10, online *n*=12) demonstrated that the standardization trainings are of great significance for the instructors regardless of the training medium. Overall, instructors trained on the online platform appreciated the quality of the standardization training content, design of the online platform such as layout, step-by-step instruction, navigational aspects, coding behind the scoring system, being free of distractors and effective use of colors. Also, they praised the practicality of the platform as it provided them with flexibility of time and place, possibility of taking breaks during the training, and being easy to use. However, several instructors emphasized that the online training should be given after receiving some face-to-face training.

Furthermore, instructors trained in face-to-face standardization praised the quality training content, selection of sample exams, videos used for the sample exams, being able to receive feedbacks from the researcher and discuss with their colleagues during the training. However, they emphasized that they should receive feedbacks after each sample exam to be able to grasp everything perfectly, and the standardization trainings should be conducted in a longer period of time instead of two weeks.

CHAPTER 5

DISCUSSION AND CONCLUSION

This study attempted to explore the effectiveness of an online standardization platform designed to train instructors for the official oral examinations. In order to get rational results, it was thought that giving standardization trainings both online and face-to-face to different groups of instructors would yield to meaningful findings as to whether online standardization trainings could prove as effective as the conventional face-to-face standardization trainings. Therefore, the researcher conducted two standardization trainings with different instructors so as to see whether instructors in both standardization mediums score sample exams consistently within their groups. As another quantitative data, the scores of instructors given to sample exams in both standardization sessions were compared to one another in order to see if there was any significant difference between the two groups. As the last quantitative data, the researcher attempted to see whether standardization trainings helped instructors score consistently with their partners in a real oral examination. Qualitative data were thought to reveal the views of instructors in both standardization trainings regarding the training content, positive aspects and concerns so that the researcher could make overall evaluations about both standardization mediums. Findings of the study were also evaluated in this chapter for further practices and studies in the future by addressing to the literature.

5.1. Discussion on Online Standardization

The findings of the study indicated that the instructors who were trained on the online standardization platform demonstrated a high level of agreement (consistency) among each other within their group, however their consistency scores were relatively lower as opposed to the face-to-face group. Along with the findings of the semi-structured interviews conducted with these instructors, the findings suggested that some instructors had difficulty in understanding how to score sample exams at the beginning of the training, and that they had questions on their minds. By addressing to the studies in the literature (Lim, 2011; Shaw, 2002; Weigle, 1998), the researcher had expected that instructors (especially the novice ones) would have lower levels of agreement with their peers at the beginning of the training. Because of this reason, the consistency analysis of the instructors was done in three parts (the first three exams, the six three exams, all ten exams) to see whether the time spent and scoring some sample exams would increase their consistency (agreement) levels with other instructors within their groups.

Even though the results of the interviews were mostly positive, there might have been other conditions regarding the lower level of agreement among instructors in the first three and six sample exams. Christmann (2017) suggested that people receiving online training could have negative outcomes due to not having enough technological literacy. The relatively low level of agreement among instructors could also arise from the lack of communication with the trainer and their colleagues during the training. During the interviews, some instructors expressed their opinions about this issue stating that it would be more beneficial for them to have feedbacks from the trainer, or combining the online training with face-to-face training would work for them better. Gill (2003) supported this idea by stating that some trainees might want to take part in a trainer-led face-to-face training so that they could exchange ideas with their peers. Elder, Barkhuizen, Knoch, and von Randow (2007) reported in their study that raters on the online platform demonstrated low levels of inter-rater agreement and concluded that the raters should have been given a face-to-face training before they embarked on the online platform, which is parallel to the quantitative findings and the analysis of the semi-structured interviews.

On the other hand, in scoring a total of ten sample exams, the instructors trained online demonstrated an excellent level of inter-rater agreement within their group. The data from semi-structured interviews suggested that the majority of instructors appreciated the practicality and design of the platform. Several instructors stated that the training content was to the point, so they were able to understand the basics of the exam, and what to do while scoring the exams. They also indicated that the flexibility of time and place of the platform was really effective for them as they could get their trainings anywhere, anytime and in any pace they wished. Upton (2006) argued that trainees could arrange their training schedules and access the platform without any time and place restrictions. Another appreciated feature of the online standardization platform was its overall design. Some instructors mentioned that they really liked the design of the platform in terms of layout, navigation, and the pacing of the training. The literature suggests that trainees might be attracted more should the training content be aided by visuals, online tutorials, and interactive content (Harasim, 2000; Khan, 2002; Powell, 2000).

5.2. Discussion on Face-to-face Standardization

The results of the study suggested that instructors who were trained in face-to-face standardization training grasped the content of the training very well and displayed a high level of consistency (agreement) within their group with regard to scoring sample exams in the training. Shohamy, Gordon, and Kraemer (1992) expressed that training could help raters produce high levels of inter-rater reliability (agreement). Furthermore, Xi and Mollaun (2011) stated that training raters helps enhance the scoring quality of raters with regard to well-understood grading criteria and make them more homogeneous thus improving intra and inter-rater reliability among raters. Both novice and experienced instructors trained face-to-face expressed their positive opinions about the training stating that they were able to learn and refresh their knowledge about the fundamentals of the exam, procedures, and rules of the exam to a great extent. Semi-structured interviews with these instructors demonstrated that they had positive feelings towards face-to-face standardization as they were able to understand the exam procedures and have discussions with the trainer and their colleagues (Arikan, 2006). Tynjälä and Häkkinen (2005) reported that adults might wish to participate in face-to-face trainings as they could share opinions and build knowledge collaboratively with their colleagues. The instructors also emphasized that receiving the training in two sessions worked well for them since it might have been too difficult for them to cover all training content in one training session. Powell

(2000) and Jung and Rha (2000) suggested that implementing face-to-face trainings are time-consuming and difficult to provide to large number of trainees. Therefore, it could be noted that the researcher might have had difficulties in implementing the standardization had it been planned for just one session.

5.3. A Comparison of Face-to-face and Online Standardization Mediums

The findings of the study suggested that both standardization mediums were favored by the instructors who were trained in them. None of the instructors mentioned that they were not able to understand the training content or they had difficulties in understanding how to score student exams. The findings for the 1st and 2nd research questions indicated that instructors in both groups showed “excellent” level of agreement (consistency) within their groups. Thus, it could be emphasized that they clearly understood the training content and how to score sample student exams. Similarly, the 3rd research question of the study attempted to find whether any significant difference between the scores of instructors in face-to-face and online standardization groups, and it was found that these two groups did not differ from one another in terms of their scores of the ten sample speaking exams. Scott, Feldman, and Underwood, (2016) emphasized that both face-to-face and online mediums of training had the similar results in terms of rater training. It was also noted in a study that instructors trained in either face-to-face training or online training were not better than the other in terms of the improvement of their rating skills. Both mediums of trained had similar effects on the instructors (Powell et al., 2010).

5.4. A comparison of Novice and Experienced Instructors in terms of Rating Experience

Weigle (1998) compared the scores of 8 novice and 8 experienced instructors as raters. The novice raters showed differences during the training sessions, however, they demonstrated similar variability in their scores in the post-test. The whole group also indicated an increased level of rating quality following the post-test. Attali (2016) found in the study that the novice and experienced instructors showed differences in their scores in the initial trainings, but in their actual rating experiences they did not differ in terms of their scoring. However, in this study, this was not the case. Both novice and experienced instructors showed differences in the consistency of scoring within their groups at different times. There were experienced and novice instructors with low consistency values in both mediums. There were also experienced and novice instructors with high consistency values in both face-to-face and online standardization mediums. Shohamy et al. (1992) argued that novice raters could produce consistent scores and there might be experienced raters with inconsistent scores as well. The findings for the 4th question also supported this situation as the assignment of instructors in the official exam was done by assigning one experienced and one novice instructors in each pair of the exam. Eight pairs of instructors in the official exam demonstrated a “substantial” agreement with their partner. This means that eight pairs of instructors including both novice and experienced instructors agreed with their partner substantially. Each of these eight pairs also included instructors from both standardization mediums.

5.5. Implications and Suggestions for Further Studies

This study might make contributions to the improvement and implementation of online standardization platforms used to train instructors for different purposes. This way, it could pave the way to the transition from conventional training to online training mediums. Thus, it might enable educational institutions to overcome problems such as time limitations (Levin, Buell, & James, 2001), and low possibility of training a number of instructors at the same time due to teaching schedules of instructors and administrative roles of trainers (Diaz & Bontenbal, 2001). Regarding the practicality of the online standardization platform, it could help institutions train their instructors more quickly and easily (Upton, 2006).

The participants in this study were not assigned to their groups (online and face-to-face standardization) after performing a descriptive analysis where their opinions were collected and analyzed prior to the sampling. Further studies might take this condition into consideration and apply a test regarding participants' preferences, technological literacy and knowledge in their fields in order to correctly assign participants into their groups so that the results might be more valid and reliable. Moreover, further studies might also focus on sampling a larger size of participants so as to cultivate more reliable and generalizable outcomes. In addition, utilization of random sampling would be a good way to satisfy the needs of generalizability, validity and reliability issues.

Lastly, the platform used for the study could only be exploited on the web environment although it was on a responsive website. The reason for this was that in the simulation of scoring sample exams, the criteria and the exam video are put next to each other so that instructors would not be distracted by scrolling down and up to see the criteria and the video. However, the platform could still be utilized on smart phones and tablets if a logical solution were found. Therefore, further studies would focus on implementing standardization training on web, smart phones, and tablets as well.

5.6. Conclusion

Overall, the online platform used to train instructors as raters for official oral examinations seemed to be as effective as its face-to-face counterpart. Instructors trained on the online standardization platform appreciated the platform's design, content, and the flexibility of their learning. Within the scope of this study, the instructors demonstrated positive feelings towards the online standardization platform for several reasons such as practicality, design, and flexibility of time and place. Therefore, it was seen that instructors had a positive tendency to receive online training in order to prepare for the actual examinations. They also had some concerns and suggestions for the platform. For this reason, it is of importance that this platform should be improved in the future. Along with the results, this study proved that instructors had been successful in their trainings and in the official examination, which means that the platform could pave the way for other studies attempting to deal with professional development issues in various professions.

REFERENCES

Anderson, N., & Henderson, M. (2005). e-PD: Blended models of sustaining teacher professional development in digital literacies. *E-Learning and digital media*, *1*(3), 383–394. https://doi.org/10.2304/elea.2004.1.3.4

Arikan, Y. D. (2006). The effects of web-supported active learning activities on teacher trainees ' attitudes towards course. *Online*, *2006*(7), 23–41.

Attali, Y. (2016). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing*. https://doi.org/10.1177/0265532215582283

Azevedo, R., Guthrie, J. T., & Seibert, D. (2005). The role of self-regulated learning in fostering students' conceptual understanding of complex systems with hypermedia. *Journal of Educational Computing Research*. https://doi.org/10.2190/dvwx-gm1t-6thq-5wc7

Babaii, E., Taghaddomi, S., & Pashmforoosh, R. (2016). Speaking self-assessment: Mismatches between learners' and teachers' criteria. *Language Testing*. https://doi.org/10.1177/0265532215590847

Bachman, L. F., & Palmer, A. S. (1996). Language testing in practice: Designing and developing useful language tests. *Oxford Applied Linguistics*. https://doi.org/10.2307/328718

Badri, M., Alnuaimi, A., Mohaidat, J., Yang, G., & Al Rashedi, A. (2016). Perception of Teachers' professional development needs, impacts, and barriers: The Abu Dhabi case. *SAGE Open*, *6*(3). https://doi.org/10.1177/2158244016662901

Baran, B., & Cagiltay, K. (2006). Teachers' experiences in online professional development environment. *Online Submission*.

Bartley, S. J., & Golek, J. H. (2004). Evaluating the cost effectiveness of online and face-to-face instruction. *Journal of Educational Technology & Society*, *7*(4), 167–175. Retrieved from http://www.jstor.org/stable/jeductechsoci.7.4.167

Bates, M. S., Phalen, L., & Moran, C. (2016). Online professional development a primer. *Phi Delta Kappan*. https://doi.org/10.1177/0031721716629662

Bauer, W. I. (2007). Research on professional development for experienced music teachers. *Journal of Music Teacher Education*. https://doi.org/10.1177/10570837070170010105

Baumgartner, E., & Hsi, S. (2019). *CILT2000 :* Synergy , technology , and teacher professional development. *11*(3), 311–315.

Bayram, İ., Altug, C., Dereli, P., Yildiz, G., & Uzun, Y. (2017). Investigating how students transfer a source text into speech through lesson study. In *European Scientific Journal* (Vol. 13). https://doi.org/10.19044/esj.2017.v13n32p49

Bennett-Levy, J., Hawkins, R., Perry, H., Cromarty, P., & Mills, J. (2012). Online cognitive behavioural therapy training for therapists: Outcomes, acceptability, and impact of support. *Australian Psychologist*. https://doi.org/10.1111/j.1742-9544.2012.00089.x

Bergman, M. (2008). Advances in mixed methods research. https://doi.org/10.4135/9780857024329

Bernard, R. M., Abrami, P. C., Borokhovski, E., Wade, C. A., Tamim, R. M., Surkes, M. A., & Bethel, E. C. (2009). A meta-analysis of three types of interaction treatments in distance education. *Review of Educational Research*, *79*(3), 1243–1289. https://doi.org/10.3102/0034654309333844

Bernstein, J., van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*. https://doi.org/10.1177/0265532210364404

Bezzina, C. (2006). Views from the trenches: Beginning teachers' perceptions about their professional development. *Journal of In-Service Education*. https://doi.org/10.1080/13674580601024515

Bhagat, K. K., Wu, L. Y., & Chang, C. Y. (2016). Development and validation of the perception of students towards online learning (POSTOL). *Educational Technology and Society*.

Bijani, H. (2018). Investigating the validity of oral assessment rater training program: A mixed-methods study of raters' perceptions and attitudes before and after training. *Cogent Education*. https://doi.org/10.1080/2331186X.2018.1460901

Bitsch, V. (2005). Qualitative research: A grounded theory example and evaluation criteria. *Journal of Agribusiness*, *23*(1,), 59612.

Bohnenkamp, J., & McMahon, J. (2001). Click, Developing an online professional development program for teachers. Technology Horizons in Education, 28 (11), 21-25.

Boisselle, L. N. (2014). Online-learning and its utility to higher education in the Anglophone Caribbean. *SAGE Open*. https://doi.org/10.1177/2158244014555118

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*. https://doi.org/10.1191/1478088706qp063oa

Brookhart, S. M. (2013). Classroom assessment in the context of motivation theory and research. In *SAGE Handbook of Research on Classroom Assessment*. https://doi.org/10.4135/9781452218649.n3

Brown, A. (2007). 3. An investigation of the rating process in the IELTS oral interview. *IELTS Collected Papers: Research in Speaking and Writing Assessment*.

Brown, A., Iwashita, N., & McNamara, T. (2005). An examination of rater orientations and test-taker performance on english-for-academic-purposes speaking tasks. *ETS Research Report Series*. https://doi.org/10.1002/j.2333-8504.2005.tb01982.x

Canaran, Ö. (2017). *A new perspective into team teaching as a continuous professional development model for English teachers.* (Unpublished doctoral dissertation). Hacettepe University, Ankara, Turkey.

Chamorro, R. (2004). Mentoring the Parentified Child: The Professional Development of the Latina(O) Psychologist. Journal of Hispanic Higher Education, 3(1), 64–72. https://doi.org/10.1177/1538192703259467

Choak, C. (2013). Asking questions: Interviews and evaluations. In *Research and Research Methods for Youth Practitioners*. https://doi.org/10.4324/9780203802571

Christmann, E. P. (2017). A comparison of the achievement of statistics students enrolled in online and face-to-face settings. *E-Learning and Digital Media*, *14*(6), 323–330. https://doi.org/10.1177/2042753017752925

Ciarocco, N. J., Dinella, L. M., Hatchard, C. J., & Valosin, J. (2016). Integrating Professional Development Across the Curriculum: An Effectiveness Study. *Teaching of Psychology*, *43*(2), 91–98. https://doi.org/10.1177/0098628316636217

Clegg, S., Hudson, A., & Steel, J. (2003). The emperor's new clothes: Globalisation and e-learning in higher education. *British Journal of Sociology of Education*. https://doi.org/10.1080/01425690301914

Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement*. https://doi.org/10.1111/j.1745-3984.2000.tb01081.x

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design & analysis issues in field settings. In *Houghton Mifflin*.

Coole, H., & Watts, M. (2013). Communal E-Learning Styles in the Online Classroom. *Research in Education*, *82*(1), 13–27. https://doi.org/10.7227/rie.82.2

Creswell, J. W. (2013). Qualitative, quantitative, and mixed methods approaches. In *Research design*.

Creswell, J. W. (2007). Qualitative enquiry & research design, choosing among five approaches. In *Book*. https://doi.org/10.1016/j.aenj.2008.02.005

Creswell, J. W., & Clark, V. L. P. (2007). Understanding mixed methods research. In

Designing and conducting mixed methods research.

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing*, *33*(1), 117–135. https://doi.org/10.1177/0265532215582282

Debowski, S., Wood, R. E., & Bandura, A. (2001). Impact of guided exploration and enactive exploration on self-regulatory mechanisms and information acquisition through electronic search. *Journal of Applied Psychology*. https://doi.org/10.1037/0021-9010.86.6.1129

Dedman, D. E., & Palmer, L. B. (2011). Field instructors and online training: An exploratory survey. *Journal of Social Work Education*, *47*(1), 151–161. https://doi.org/10.5175/JSWE.2011.200900057

DeLoose, S., Unger, K. L., Zhang, L., & Moseley, J. L. (2009). Moodling around: A tool for interactive technologies. *Educational Technology*, *49*(5), 28–32. Retrieved from http://jproxy.lib.ecu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=ehh&AN=44046437&site=ehost-live

Denzin, N. K., & Lincoln, Y. S. (2005). *The SAGE Handbook of Qualitative Research*. Retrieved from https://books.google.com.tr/books?id=X85J8ipMpZEC

Desimone, L. M. (2009). Improving Impact Studies of Teachers' Professional Development: Toward Better Conceptualizations and Measures. *Educational Researcher*, *38*(3), 181–199. https://doi.org/10.3102/0013189X08331140

Dewhurst, D. G., MacLeod, H. A., & Norris, T. A. M. (2000). Independent student learning aided by computers: An acceptable alternative to lectures? *Computers* *and Education*. https://doi.org/10.1016/S0360-1315(00)00033-6

Diaz, D. P., & Bontenbal, K. F. (2001). Learner Preferences : Developing a Learner-Centered Environment in the Online or Mediated Classroom. *Distance Education*.

Drouin, M. A. (2010). Group-Based Formative Summative Assessment Relates to Improved Student Performance and Satisfaction. *Teaching of Psychology*, *37*(2), 114–118. https://doi.org/10.1080/00986281003626706

Duffy, C. (2016). ICON: Radical professional development in the conservatoire. *Arts and Humanities in Higher Education*, *15*(3–4), 376–385. https://doi.org/10.1177/1474022216647385

Earley, P., & Bubb, S. (2004). Leading and managing continuing professional development: Developing people, developing schools. In *Leading and Managing Continuing Professional Development: Developing People, Developing Schools*. https://doi.org/10.4135/9781446279601

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*. https://doi.org/10.1177/0265532207086780

Edinger, M. J. (2017). Online Teacher Professional Development for Gifted Education: Examining the Impact of a New Pedagogical Model. *Gifted Child Quarterly*, *61*(4), 300–312. https://doi.org/10.1177/0016986217722616

Edwards, C. M., Rule, A. C., & Boody, R. M. (2017). Middle School Students' Mathematics Knowledge Retention. *Journal of Educational Technology & Society*, *20*(4), 1–10. Retrieved from http://www.jstor.org/stable/26229200

Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing*. https://doi.org/10.1177/0265532207071511

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual Feedback to Enhance Rater Training: Does It Work? *Language Assessment Quarterly*. https://doi.org/10.1207/s15434311laq0203_1

Ellis, N. C., Römer, U., & O'Donnell, M. B. (2016). Usage-based approaches to language acquisition and processing: Cognitive and corpus investigations of construction grammar. *Language Learning Monograph Series*. https://doi.org/10.1111/lang.12177

Enriquez, A. G. (2010). Enhancing Student Performance Using Tablet Computers. *College Teaching*. https://doi.org/10.1080/87567550903263859

Erickson, A. S. G., Noonan, P. M., & Mccall, Z. (2017). Effectiveness of Online Professional Development for Rural Special Educators. *Rural Special Education Quarterly*, *31*(1), 22–32. https://doi.org/10.1177/875687051203100104

Erlam, R., von Randow, J., & Read, J. (2013). Investigating an online rater training program: Product and process. *Papers in Language Testing and Assessment*.

Eros, J. (2013). Second-stage music teachers' perceptions of their professional development. *Journal of Music Teacher Education*, *22*(2), 20–33. https://doi.org/10.1177/1057083712438771

Ertmer, P. A. (2005). Teacher pedagogical beliefs: The final frontier in our quest for technology integration? *Educational Technology Research and Development*.

https://doi.org/10.1007/BF02504683

Etikan, I., Musa, S. A., & Alkassim, R. S. (2014). A comparison of convenience sampling and purposive sampling. *Journal of Nursing*. https://doi.org/10.6224/JN.61.3.105

Eun, B. (2018). Adopting a stance: Bandura and Vygotsky on professional development. *Research in Education*. https://doi.org/10.1177/0034523718793431

Fahim, M., & Bijani, H. (2011). The Effects of Rater Training on Raters' Severity and Bias in Second Language Writing Assessment. *Iranian Journal of Language Testing*, *1*.

Faibisoff, S. G., & Willis, D. J. (2010). Distance Education: Definition and Overview. *Journal of Education for Library and Information Science*. https://doi.org/10.2307/40323650

Fazlollahtabar, H., & Yousefpoor, N. (2009). Cost optimization in e-learning-based education systems: Implementation and learning sequence. *E-Learning*, *6*(2), 198–205. https://doi.org/10.2304/elea.2009.6.2.198

Fisher, J. B., Schumaker, J. B., Culbertson, J., & Deshler, D. D. (2010). Effects of a Computerized Professional Development Program on Teacher and Student Outcomes. *Journal of Teacher Education*. https://doi.org/10.1177/0022487110369556

Fleischmann, K. (2018). Online design education: Searching for a middle ground. *Arts and Humanities in Higher Education*, 147402221875823. https://doi.org/10.1177/1474022218758231

Flowers, L. O., White, E. N., Raynor, J. E., & Bhattacharya, S. (2012). African American students' participation in online distance education in STEM disciplines: Implications for HBCUs. *SAGE Open*, *2*(2), 1–5. https://doi.org/10.1177/2158244012443544

Fox, J. C. (2015). *The Ultimate Guide to Excellent Teaching and Training: Face-To-Face and Online*. Retrieved from https://books.google.com.tr/books?id=MwYvCgAAQBAJ

Fulcher, G. (2014). Testing Second Language Speaking. In *Testing Second Language Speaking*. https://doi.org/10.4324/9781315837376

Gardner, H. (2011). Frames of Mind: The Theory of Multiple Intelligences. In *the theory of multiple intelligences*. https://doi.org/10.2307/3324261

Gerard, L. F., Varma, K., Corliss, S. B., & Linn, M. C. (2011). Professional Development for Technology-Enhanced Inquiry Science. *Review of Educational Research*, *81*(3), 408–448. https://doi.org/10.3102/0034654311415121

Giguere, P., & Minotti, J. (2003). Developing High-Quality Web-Based Training for Adult Learners. *Educational Technology*, *43*(4), 57–58. Retrieved from http://jproxy.lib.ecu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=ehh&AN=21652313&site=ehost-live

Gill, S. (2003). Myths and Reality of e-Learning. *Educational Technology*, *43*(1), 20–24.

Glomb, N., Midenhall, T., Mason, L. L., & Salzberg, C. (2017). Reducing Isolation through Regional Mentors and Learning Communities: A Way to Support Rural

Learners. *Rural Special Education Quarterly*. https://doi.org/10.1177/875687050902800405

Gradel, K., & Edson, A. J. (2012). Integrating Cloud-Based Strategies and Tools in Face-to-Face Training Sessions to Increase the Impact of Professional Development. *Journal of Educational Technology Systems*. https://doi.org/10.2190/et.40.2.c

Gribbons, B., & Herman, J. (1997). True and Quasi-Experimental Designs. *Practical Assessment, Research & Evaluation*.

Griffin, C. C., Dana, N. F., Pape, S. J., Algina, J., Bae, J., Prosser, S. K., & League, M. B. (2017). Prime Online : Exploring Teacher Professional Development for Creating Inclusive Elementary Mathematics Classrooms . *Teacher Education and Special Education: The Journal of the Teacher Education Division of the Council for Exceptional Children*. https://doi.org/10.1177/0888406417740702

Guskey, T. R. (2002). Professional development and teacher change. *Teachers and Teaching: Theory and Practice*. https://doi.org/10.1080/135406002100000512

Hamilton, J., Reddel, S., & Spratt, M. (2001). Teachers' perceptions of on-line rater training and monitoring. *System*. https://doi.org/10.1016/S0346-251X(01)00036-7

Hamp-Lyons, L. (2012). Writing teachers as assessors of writing. In *Exploring the Dynamics of Second Language Writing*. https://doi.org/10.1017/cbo9781139524810.012

Haney, J. J., & Lumpe, A. T. (1995). A Teacher Professional Development Framework Guided by Reform Policies, Teachers' Needs, and Research. *Journal* *of Science Teacher Education*, *6*(4), 187–196. Retrieved from http://www.jstor.org/stable/43156047

Harasim, L. (2000). Shift happens: Online education as a new paradigm in learning. *Internet and Higher Education*. https://doi.org/10.1016/S1096-7516(00)00032-4

Hardy, I. (2016). *Teacher Professional Development : A Sociological Study of Senior Educators ' PD Priorities in Ontario Ian Hardy*. *32*(3), 509–532.

Hartley, J. (2002). Studying for the Future. *Journal of Further and Higher Education*. https://doi.org/10.1080/03098770220149576

Hattie, J. (2008). Visible learning: A synthesis of over 800 meta-analyses relating to achievement. In *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*. https://doi.org/10.4324/9780203887332

Hill, H. C. (2007). Learning in the teaching workforce. *Future of Children*. https://doi.org/10.1353/foc.2007.0004

Hill, H. C., Beisiegel, M., & Jacob, R. (2013). Professional Development Research. *Educational Researcher*, *42*(9), 476–487. https://doi.org/10.3102/0013189x13512674

Holden, J. T., & Westfall, P. J.-L. (2010). An instructional media selection guide for distance learning - implications for blended learning. *Learning*.

Holmes, A., Polhemus, L., & Jennings, S. (2005). Catie: A Blended Approach to Situated Professional Development. *Journal of Educational Computing Research*, *32*(4), 381–394. https://doi.org/10.2190/f97w-quj4-g7yg-fpxc

Hoover, N. R., & Abrams, L. M. (2013). Teachers' Instructional Use of Summative Student Assessment Data. *Applied Measurement in Education*. https://doi.org/10.1080/08957347.2013.793187

Hou, S. H., Horng, R. Y., & Chen, P. H. (2016a). Professional Development in Practice. *Comprehensive Psychology*. https://doi.org/10.1177/2165222816656497

Hou, S. H., Horng, R. Y., & Chen, P. H. (2016b). Professional Development in Practice. *Comprehensive Psychology*, *5*, 216522281665649. https://doi.org/10.1177/2165222816656497

Houle, J. C. (2006). Professional development for urban principals in underperforming schools. *Education and Urban Society*, *38*(2), 142–159. https://doi.org/10.1177/0013124505282611

Howard, R., & McGrath, I. (1995). *Distance education for language teachers: A UK perspective*. Multilingual matters.

Huai, N., Braden, J. P., White, J. L., & Elliott, S. N. (2009). Effect of an Internet-Based Professional Development Program on Teachers' Assessment Literacy for All Students. *Teacher Education and Special Education: The Journal of the Teacher Education Division of the Council for Exceptional Children*, *29*(4), 244–260. https://doi.org/10.1177/088840640602900405

Huang, H. T. D., Hung, S. T. A., & Plakans, L. (2018). Topical knowledge in L2 speaking assessment: Comparing independent and integrated speaking test tasks. *Language Testing*. https://doi.org/10.1177/0265532216677106

Hughes, J., Morrison, L., & Dobos, L. (2018). Re-making teacher professional development. *Studies in Health Technology and Informatics*. https://doi.org/10.3233/978-1-61499-923-2-602

Hur, J. W., & Hara, N. (2007). Factors Cultivating Sustainable Online Communities for K-12 Teacher Professional Development. *Journal of Educational Computing Research, 36*(3), 245–268. https://doi.org/10.2190/37h8-7gu7-5704-k470

İlhan, M., & Cetin, B. (2014). Performans Değerlendirmeye Karışan Puanlayıcı Etkilerini Azaltmanın Yollarından Biri Olarak Puanlayıcı Eğitimleri: Kuramsal Bir Analiz. *Journal of European Education; Vol 4 No 2 (2014)*. Retrieved from http://www.eu-journal.org/index.php/JEE/article/view/205

In'nami, Y., & Koizumi, R. (2016). Task and rater effects in L2 speaking and writing: A synthesis of generalizability studies. *Language Testing*. https://doi.org/10.1177/0265532215587390

Isaacs, T., & Harding, L. (2017). Pronunciation assessment. *Language Teaching*. https://doi.org/10.1017/S0261444817000118

Isaacs, T., Trofimovich, P., & Foote, J. A. (2018). Developing a user-oriented second language comprehensibility scale for English-medium universities. *Language Testing*. https://doi.org/10.1177/0265532217703433

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*. https://doi.org/10.1093/applin/amm017

James, G. (2002). Advantages and disadvantages of online learning.

Jao, L., & McDougall, D. (2015). The collaborative teacher inquiry project: A purposeful professional development initiative. *Canadian Journal of Education*.

Johnson, W. W. (2014). How to Be a Successful Teacher of Professional Development. *Journal of Contemporary Criminal Justice*, *30*(4), 443–454. https://doi.org/10.1177/1043986214541609

Jonassen, D. H., Peck, K. L., & Wilson, B. G. (1999). Learning with technology: A constructivist perspective. In *Special Education*.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*. https://doi.org/10.1016/j.edurev.2007.05.002

Jung, I., & Rha, I. (2000). Effectiveness and Cost-Effectiveness of Online Education: A Review of the Literature. *Educational Technology*, *40*(4), 57–60. Retrieved from http://jproxy.lib.ecu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=ehh&AN=21606665&site=ehost-live

Kang, O., Rubin, D., & Kermad, A. (2019). The effect of training and rater differences on oral proficiency assessment. *Language Testing*, 0265532219849522. https://doi.org/10.1177/0265532219849522

Kao, C. P., Tsai, C. C., & Shih, M. (2014). Development of a survey to measure self-efficacy and attitudes toward web-based professional development among elementary school teachers. *Educational Technology and Society*.

Keis, O., Grab, C., Schneider, A., & Öchsner, W. (2017). Online or face-to-face instruction? A qualitative study on the electrocardiogram course at the University of Ulm to examine why students choose a particular format. *BMC Medical*

Education. https://doi.org/10.1186/s12909-017-1053-6

Keller, B. (2005). Teachers Flocking to Online Sources To Advance And Acquire Knowledge. *Education Week*.

Kennedy, M. M. (2016). How Does Professional Development Improve Teaching? *Review of Educational Research*, *86*(4), 945–980. https://doi.org/10.3102/0034654315626800

Khan, B. H. (2002). Dimensions of E-Learning. *Educational Technology*, *42*(1), 59–60. Retrieved from http://www.jstor.org/stable/44428725

Kirschner, P. A., Jochems, W. M. G., & Kreijns, K. (2005). Is Technology-Based Collaborative Learning Antisocial? or, What We Are Doing to Make It So! *Educational Technology*, *45*(5), 8–12.

Kitzes, J. A., Kalishman, S., Kingsley, D. D., Mines, J., & Lawrence, E. (2009). Palliative medicine death rounds: Small group learning on a vital subject. *American Journal of Hospice and Palliative Medicine*, *25*(6), 483–491. https://doi.org/10.1177/1049909108322296

Kleiman, G. (2004). *Meeting the Need for High Quality Teachers: e-Learning Solutions*.

Kleinsasser, R. C., & Silverman, D. (2006). Interpreting Qualitative Data Methods for Analysing Talk, Text and Interaction. *The Modern Language Journal*. https://doi.org/10.2307/329190

Knoch, U., Fairbairn, J., & Huisman, A. (2016). *An evaluation of an online rater training program for the speaking and writing sub-tests of the Aptis test*. *5*(Papers

In Language Testing And Assessment), 90–106. Retrieved from http://hdl.handle.net/11343/115276

Kondo, Y. (2010). Examination of rater training effect and rater eligibility. *Pan-Pacific Association of Applied Linguistics*.

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*. https://doi.org/10.1016/j.jcm.2016.02.012

Kottner, J., Audige, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., Streiner, D. L. (2011). Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *International Journal of Nursing Studies*. https://doi.org/10.1016/j.ijnurstu.2011.01.016

Krefting, L. (1991). Rigor in qualitative research: the assessment of trustworthiness. *The American Journal of Occupational Therapy. : Official Publication of the American Occupational Therapy Association*. https://doi.org/10.5014/ajot.45.3.214

Kunnavatana, S. S., Bloom, S. E., Samaha, A. L., & Dayton, E. (2013). Training Teachers to Conduct Trial-Based Functional Analyses. *Behavior Modification*. https://doi.org/10.1177/0145445513490950

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing*, *17*(1), 1-42. https://doi.org/10.1177/026553220001700101

Larreamendy-Joerns, J., & Leinhardt, G. (2006). Going the Distance With Online Education. *Review of Educational Research*, *76*(4), 567–605.

https://doi.org/10.3102/00346543076004567

Larson, J. W. (2019). Language Learning : Using the Proven and Proving the New. *Testing Oral Language Skills via the Computer*. *18*(1), 53–66.

Latchem, C., H. Odabasi, F., & Kabakci, I. (2006). Online Professional Development for University Teaching in Turkey: A Proposal. In *Online Submission* (Vol. 5).

Leung, C., & Lewkowicz, J. (2010). Expanding Horizons and Unresolved Conundrums: Language Testing and Assessment. *TESOL Quarterly*, *40*(1), 211. https://doi.org/10.2307/40264517

Levin, S. R., Waddoups, G. L., Levin, J., & Buell, J. (2001). Highly Interactive and Effective Online Learning Environments for Teacher Professional Development. *International Journal of Educational Technology*, *2*(2). Retrieved from https://www.learntechlib.org/p/92912

Lim, C. P. (2002). Trends in Online Learning and Their Implications for Schools. *Educational Technology*, *42*(6), 43–48. Retrieved from http://jproxy.lib.ecu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=ehh&AN=21539544&site=ehost-live

Lim, D. H. (2002). Perceived Differences between Classroom and Distance Education: Seeking Instructional Strategies for Learning Applications. *International Journal of Educational Technology*.

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*. https://doi.org/10.1177/0265532211406422

Lin, A. H., Chiu, H., Lin, H., & Chiu, H. (2019). *Using Computers to Support a Beginning Teacher ' s Professional Development Using Computers to Support a Beginning Teacher ' s Professional Development*. *9*(4), 367–373.

Lin, D., & Liu, S. (2018). Assessment in Second Language Pronunciation. By O. Kang and A. Ginther (Eds.), London and New York: Routledge, 2018, 190 pp., $39.91 (paperback). ISBN: 978-1-138-85687-5 (pbk). *Studies in English Language Teaching*. https://doi.org/10.22158/selt.v6n2p155

Lincoln, Y. S., & Guba, E. G. (1985). Establishing Trustworthiness. In *Naturalistic Inquiry*.

Ling, G., Mollaun, P., & Xi, X. (2014). A study on the impact of fatigue on human raters when scoring speaking responses. *Language Testing*, *31*(4), 479–499. https://doi.org/10.1177/0265532214530699

Loeb, S., Miller, L. C., & Strunk, K. O. (2009). The State Role in Teacher Professional Development and Education Throughout Teachers' Careers. *Education Finance and Policy*, *4*(2), 212–228. https://doi.org/10.1162/edfp.2009.4.2.212

Lorraine M. B. (2016). Formative Assessment at Work in the Classroom. *The Mathematics Teacher*, *110*(1), 46. https://doi.org/10.5951/mathteacher.110.1.0046

Lui, C. J., Ferrin, S. E., Baum, D. R., & Randall, V. E. (2018). The Preferred Perceptual Learning Styles of Hispanic Higher Education Students. *Journal of Hispanic Higher Education*. https://doi.org/10.1177/1538192718801793

Lumley, T. (2005). *Assessing Second Language Writing: The Rater's Perspective*.

Lumley, T., & Mcnamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*. https://doi.org/10.1177/026553229501200104

Lumley, T., & O'Sullivan, B. (2005). The effect of test-taker gender, audience and topic on task performance in tape-mediated assessment of speaking. *Language Testing*. https://doi.org/10.1191/0265532205lt303oa

Lunz, M. E., & Stahl, J. A. (1990). Judge Consistency and Severity Across Grading Periods. *Evaluation & the Health Professions*. https://doi.org/10.1177/016327879001300405

Luoma, S. (2004). Speaking Scales. *Assessing Speaking*. https://doi.org/10.1017/CBO9780511733017

Masters, J., De Kramer, R. M., O'Dwyer, L. M., Dash, S., & Russell, M. (2010). The Effects of Online Professional Development on Fourth Grade English Language Arts Teachers' Knowledge and Instructional Practices. *Journal of Educational Computing Research*, *43*(3), 355–375. https://doi.org/10.2190/ec.43.3.e

May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing*. https://doi.org/10.1177/0265532209104668

Mayadas, A. F., Bourne, J., & Bacsich, P. (2019). *J63JTO*. *323*(5910), 85–89.

Mayer, R. E., & Mayer, R. E. (2012). A Cognitive Theory of Multimedia Learning. In *Multimedia Learning*. https://doi.org/10.1017/cbo9781139164603.004

McHugh, M. L. (2012). Lessons in biostatistics Interrater reliability : the kappa statistic. *Biochemica Medica*.

McMillen, J. C., Hawley, K. M., & Proctor, E. K. (2016). Mental Health Clinicians' Participation in Web-Based Training for an Evidence Supported Intervention: Signs of Encouragement and Trouble Ahead. *Administration and Policy in Mental Health and Mental Health Services Research*. https://doi.org/10.1007/s10488-015-0645-x

Means, B., Toyama, Y., Murphy, R., Bakia, M., Jones, K., U.S. Department of Education, O. of P., & Development, E. and P. (2009). Evaluation of evidence-based practices in online learning. In *Structure*.

Melber, L. M., Cox-Petersen, A. M., Berg, C., & Enochs, L. (2005). Teacher Professional Development and Informal Learning Environments: Investigating Partnerships and Possibilities. *Journal of Science Teacher Education*, *16*(2), 103–120. Retrieved from http://www.jstor.org/stable/43156359

Merriam, S. B. (1998). Qualitative research and case study applications in education. In *Dados*.

Mixon, C. S. (2017). *Evaluating the Impact of Online Professional Development on Teachers' Use of a Targeted Behavioral Classroom Intervention*. Retrieved from http://rave.ohiolink.edu/etdc/view?acc_num=ohiou1493932450903862

Moreno, R., & Mayer, R. E. (1999). Cognitive principles of multimedia learning: The role of modality and contiguity. *Journal of Educational Psychology*. https://doi.org/10.1037/0022-0663.91.2.358

Murphy, G. A., & Calway, B. A. (2010a). Enterprise Professional Development – Evaluating Learning. *Industry and Higher Education*. https://doi.org/10.5367/ihe.2010.0008

Murphy, G. A., & Calway, B. A. (2010b). Enterprise Professional Development – Evaluating Learning. *Industry and Higher Education*, *24*(5), 343–359. https://doi.org/10.5367/ihe.2010.0008

Murphy, J. M. (2006). Oral Communication in TESOL: Integrating Speaking, Listening, and Pronunciation. *TESOL Quarterly*, *25*(1), 51. https://doi.org/10.2307/3587028

Myers, C. B., Bennett, D., Brown, G., Henderson, T., Journal, S., January, N., … Henderson, T. (2004). *International Forum of Educational Technology & Society Emerging Online Learning Environments and Student Learning : An Analysis of Faculty Perceptions Published by : International Forum of Educational Technology & Society Linked references are available* . *7*(1), 77–86.

Myford, C. M., & Wolfe, E. W. (2000). Monitoring sources of variability within the test of Spoken English Assessment System. *ETS Research Report Series*. https://doi.org/10.1002/j.2333-8504.2000.tb01829.x

Nolen, S. B. (2011). *Engaging Students in Learning: A Special Issue Dedicated to Jere Brophy*. *50*(4), 319–326. https://doi.org/10.1080/00405841.201

O'sullivan, B., Weir, C. J., & Saville, N. (2002). Using observation checklists to validate speaking-test tasks. *Language Testing*. https://doi.org/10.1191/0265532202lt219oa

Onwuegbuzie, A. J., & Leech, N. L. (2007). Validity and qualitative research: An oxymoron? *Quality and Quantity*. https://doi.org/10.1007/s11135-006-9000-3

Opfer, V. D., & Pedder, D. (2011). Conceptualizing Teacher Professional Learning

Published by : American Educational Research Association Stable URL : http://www.jstor.org/stable/23014297 Conceptualizing Teacher Professional Learnin. *Review of Educational Research*, *81*(3), 376–407.

Paas, F., Renkl, A., & Sweller, J. (2004). Cognitive Load Theory and Instructional Design: Recent Developments. *Educational Psychologist*. https://doi.org/10.1207/s15326985ep3801_1

Parkes, J., Zimmaro, D., Parkes, J., & Zimmaro, D. (2018). Formative and Summative Assessments. In *The College Classroom Assessment Compendium*. https://doi.org/10.4324/9781315283852-25

Patton, K., Parker, M., & Tannehill, D. (2015). Helping Teachers Help Themselves: Professional Development That Makes a Difference. *NASSP Bulletin*, *99*(1), 26–42. https://doi.org/10.1177/0192636515576040

Patton, M. Q. (2014). *Qualitative Research & Evaluation Methods: Integrating Theory and Practice*. Retrieved from https://books.google.com.tr/books?id=dCNhngEACAAJ

Patton, M. Q. (2002). Qualitative research and evaluation methods. In *Qualitative Inquiry*. https://doi.org/10.2307/330063

Pella, S. (2015). Pedagogical Reasoning and Action : Affordances of Practice-Based. *Teacher Education Quarterly*, *42*(3), 81–101.

Penuel, W. R., Fishman, B. J., Yamaguchi, R., & Gallagher, L. P. (2007). What Makes Professional Development Effective? Strategies That Foster Curriculum Implementation. *American Educational Research Journal*, *44*(4), 921–958. https://doi.org/10.3102/0002831207308221

Peters, O. (2003). Learning With New Media in Distance Education. In *Handbook of Distance Education*.

Peterson, C. L., & Bond, N. (2004). Online compared to face-to-face teacher preparation for learning standards-based planning skills. *Journal of Research on Technology in Education*. https://doi.org/10.1080/15391523.2004.10782419

Poekert, P. (2011). The pedagogy of facilitation: Teacher inquiry as professional development in a Florida elementary school. *Professional Development in Education*. https://doi.org/10.1080/19415251003737309

Portney, L. G., & Watkins, M. P. (2000). Statistical measures of reliability. In *Foundations of clinical research : applications to practice*.

Powell, D. R., Diamond, K. E., Burchinal, M. R., & Koehler, M. J. (2010). Effects of an Early Literacy Professional Development Intervention on Head Start Teachers and Children. *Journal of Educational Psychology*. https://doi.org/10.1037/a0017763

Powell, G. C. (2000). Are You Ready for Web-Based Training? *Educational Technology*, *40*(3), 52–55. Retrieved from http://www.jstor.org/stable/44428603

Putnam, R. T., & Borko, H. (2000). What Do New Views of Knowledge and Thinking Say About Learning ? *American Educational Researcher*.

Qu, W., & Zhang, C. (2013). The Analysis of Summative Assessment and Formative Assessment and Their Roles in College English Assessment System. *Journal of Language Teaching and Research*. https://doi.org/10.4304/jltr.4.2.335-339

Ragan, K. P. (2017). *Online Versus Face-to-Face Course Learning Effectiveness : Measured Outcomes for Intermediate Financial Management*. *43*(2), 243–261.

Remler, D. K., & Van Ryzin, G. G. (2011). Research methods in practice: Strategies for description and causation. In *Research methods in practice: Strategies for description and causation.*

Rohner, R. P., & Katz, L. (2004). Testing for Validity and Reliability in Cross-Cultural Research. *American Anthropologist*. https://doi.org/10.1525/aa.1970.72.5.02a00060

Römer, U. (2017). Language assessment and the inseparability of lexis and grammar: Focus on the construct of speaking. *Language Testing*. https://doi.org/10.1177/0265532217711431

Russell, J. D., & Blake, B. L. (1988). Formative and Summative Evaluation of Instructional Products and Learners. *Educational Technology, 28*(9), 22–28.

Saka, Y. (2013). Who are the Science Teachers that Seek Professional Development in Research Experience for Teachers (RET's)? Implications for Teacher Professional Development. *Journal of Science Education and Technology, 22*(6), 934–951. Retrieved from http://www.jstor.org/stable/24019768

Salehi, A., Strawderman, L., Huang, Y., Ahmed, S., & Babski-Reeves, K. (2010). Effectiveness of Three Training Delivery Methods in a Voluntary Program. *Human Factors and Ergonomics Society Annual Meeting Proceedings*. https://doi.org/10.1518/107118109x12524444844802

Sales, G. C., & Al-Rahbi, F. (2008). Building Capacity for Oman's Online Teacher Training: Making an International Partnership Work. *Educational Technology*, *48*(2), 34–37.

Scagnoli, N. (2009). A Review of Online Learning. *Policy Futures in Education*, *7*(5), 555–565.

Schaaf, D. N. (2018). Assistive Technology Instruction in Teacher Professional Development. *Journal of Special Education Technology*. https://doi.org/10.1177/0162643417753561

Schwandt, T. A., Lincoln, Y. S., & Guba, E. G. (2007). Judging interpretations: But is it rigorous? Trustworthiness and authenticity in naturalistic evaluation. *New Directions for Evaluation*. https://doi.org/10.1002/ev.223

Scott, M., Feldman, B. N., & Underwood, M. (2016). Delivering Professional Development in Suicide Prevention. *Pedagogy in Health Promotion*, *2*(4), 266–275. https://doi.org/10.1177/2373379916658667

Selwyn, N., Gorard, S., & Furlong, J. (2005). Adult learning in the digital age: Information technology and the learning society. In *Adult Learning in the Digital Age: Information Technology and the Learning Society*. https://doi.org/10.4324/9780203003039

Semmar, Y. (2013). Distance Learners and Academic Achievement: The Roles of Self-Efficacy, Self-Regulation and Motivation. *Journal of Adult and Continuing Education*. https://doi.org/10.7227/jace.12.2.9

Sharkey, N. S., & Murnane, R. J. (2006). Tough Choices in Designing a Formative Assessment System. *American Journal of Education*, *112*(4), 572–588. https://doi.org/10.1086/505060

Sharma. (2017). Pros and cons of different sampling techniques. *International Journal of Applied Research*.

Shaw, S. (2002). The effect of training and standardisation on rater judgement and inter-rater reliability. *Research Notes*. https://doi.org/10.1007/s10439-013-0884-5

Shenton, A. K. (2004). Strategies for ensuring trustworthiness in qualitative research projects. *Education for Information*. https://doi.org/10.3233/EFI-2004-22201

Shirley, M., & Irving, K. (2015). Connected Classroom Technology Facilitates Multiple Components of Formative...: EBSCOhost. *J Sci Educ Technol*, *24*(1), 56–68.

Shohamy, E. (2006). Affective Considerations in Language Testing. *The Modern Language Journal*, *66*(1), 13. https://doi.org/10.2307/327810

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The Effect of Raters' Background and Training on the Reliability of Direct Writing Tests. *The Modern Language Journal*. https://doi.org/10.1111/j.1540-4781.1992.tb02574.x

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*. https://doi.org/10.1037/0033-2909.86.2.420

Sinclair, P., Fitzgerald, J. E. F., Hornby, S. T., & Shalhoub, J. (2015). Mentorship in surgical training: Current status and a needs assessment for future mentoring programs in surgery. *World Journal of Surgery*. https://doi.org/10.1007/s00268-014-2774-x

Skylar, A. A., Higgins, K., Boone, R., Jones, P., Pierce, T., & Gelfer, J. (2017).

Distance Education: An Exploration of Alternative Methods and Types of Instructional Media in Teacher Education. *Journal of Special Education Technology*. https://doi.org/10.1177/016264340502000303

Steward, B. ., Mickelson, S. ., & Brumm, T. . (2004). Formative and Summative Assessment Techniques for Continuous Agricultural Technology Classroom Improvement. *NACTA Journal*, *48*(8), 33–41.

Sun, M., Penuel, W. R., Frank, K. A., Gallagher, H. A., & Youngs, P. (2013). Shaping Professional Development to Promote the Diffusion of Instructional Expertise Among Teachers. *Educational Evaluation and Policy Analysis*, *35*(3), 344–369. https://doi.org/10.3102/0162373713482763

Sunal, D. W., Hodges, J., Sunal, C. S., Whitaker, K. W., Freeman, L. M., Edwards, L., … Odell, M. (2001). Teaching Science in Higher Education: Faculty Professional Development and Barriers to Change. *School Science and Mathematics*. https://doi.org/10.1111/j.1949-8594.2001.tb18027.x

Tajeddin, Z., & Alemi, M. (2014). Pragmatic Rater Training: Does It Affect Non-native L2 Teachers' Rating Accuracy and Bias? *Iranian Journal of Language Testing*, *4*, 66–83.

Tajeddin, Z., Alemi, M., & Pashmforoosh, R. (2011). Non-native teachers' rating criteria for L2 speaking: Does a rater training program make a difference? *TELL*, *5*, 125–153.

Takir, A., & Aksu, M. (2012). The Effect of an Instruction Designed by Cognitive Load Theory Principles on 7th Grade Students' Achievement in Algebra Topics and Cognitive Load. *Creative Education*. https://doi.org/10.4236/ce.2012.32037

Terry, R. M., & Hughes, A. (2006). Testing for Language Teachers. *The Modern Language Journal*. https://doi.org/10.2307/327632

The Web-Based Education Commission. (2000). Power of the Internet for Learning: Moving from Promise to Practice. *Journal of Government Information*. https://doi.org/10.1016/S1352-0237(00)00192-1

Tondeur, J., Forkosh-Baruch, A., Prestridge, S., Albion, P., & Edirisinghe, S. (2016). Responding to Challenges in Teacher Professional Development for ICT Integration in Education. *Journal of Educational Technology & Society*, *19*(3), 110–120. Retrieved from http://www.jstor.org/stable/jeductechsoci.19.3.110

Torff, B., Sessions, D., & Byrnes, K. (2005). Assessment of teachers' attitudes about professional development. *Educational and Psychological Measurement*, *65*(5), 914–924. https://doi.org/10.1177/0013164405275664

Tynjälä, P. T., & Häkkinen, P. H. (2005). E-learning at work: Theoretical underpinnings and pedagogical challenges. *Journal of Workplace Learning*. https://doi.org/10.1108/13665620510606742

Upton, D. (2006). Online learning in speech and language therapy: Student performance and attitudes. *Education for Health: Change in Learning and Practice*. https://doi.org/10.1080/13576280500534735

Van Driel, J. H., & Berry, A. (2012). Teacher Professional Development Focusing on Pedagogical Content Knowledge. *Educational Researcher*, *41*(1), 26–28. https://doi.org/10.3102/0013189x11431010

Venkatesh, V., & Davis, F. D. (2000). A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. *Management Science*.

https://doi.org/10.1287/mnsc.46.2.186.11926

Villar, L. M., & Alegre, O. M. (2007). An Innovative Junior Faculty Online Development Programme. *E-Learning and Digital Media*, *3*(4), 599–612. https://doi.org/10.2304/elea.2006.3.4.599

Volante, L., & Beckett, D. (2011). Formative assessment and the contemporary classroom: Synergies and tensions between research and practice. *Canadian Journal of Education*.

Wang, B. (2014). On Rater Agreement and Rater Training. *English Language Teaching*. https://doi.org/10.5539/elt.v3n1p108

Wayne, A. J., Yoon, K. S., Zhu, P., Cronen, S., & Garet, M. S. (2008). Experimenting With Teacher Professional Development: Motives and Methods. *Educational Researcher*, *37*(8), 469–479. https://doi.org/10.3102/0013189x08327154

Webster-Wright, A. (2009). Reframing Professional Development Through Understanding Authentic Professional Learning. *Review of Educational Research*, *79*(2), 702–739. https://doi.org/10.3102/0034654308330970

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*. https://doi.org/10.1177/026553229401100206

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*. https://doi.org/10.1177/026553229801500205

Weir, C. J. (2005). Language Testing and Validation: An Evidence-Based Approach by WEIR, CYRIL J. *The Modern Language Journal*.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*. https://doi.org/10.1177/026553229301000306

Wind, S. A. (2019). Examining the Impacts of Rater Effects in Performance Assessments. *Applied Psychological Measurement*. https://doi.org/10.1177/0146621618789391

Xi, X., & Mollaun, P. (2011). Using raters from India to score a large-scale speaking test. *Language Learning*. https://doi.org/10.1111/j.1467-9922.2011.00667.x

Yan, X. (2014a). An examination of rater performance on a local oral English proficiency test: A mixed-methods approach. *Language Testing*, *31*(4), 501–527. https://doi.org/10.1177/0265532214536171

Yan, X. (2014b). An examination of rater performance on a local oral English proficiency test: A mixed-methods approach. *Language Testing*. https://doi.org/10.1177/0265532214536171

Yilmaz, K. (2013). Comparison of quantitative and qualitative research traditions: Epistemological, theoretical, and methodological differences. *European Journal of Education*. https://doi.org/10.1111/ejed.12014

Zhang, Y., & Elder, C. (2011). Judgments of oral proficiency by non-native and native english speaking teacher raters: Competing or complementary constructs? *Language Testing*, *28*(1), 31–50. https://doi.org/10.1177/0265532209360671

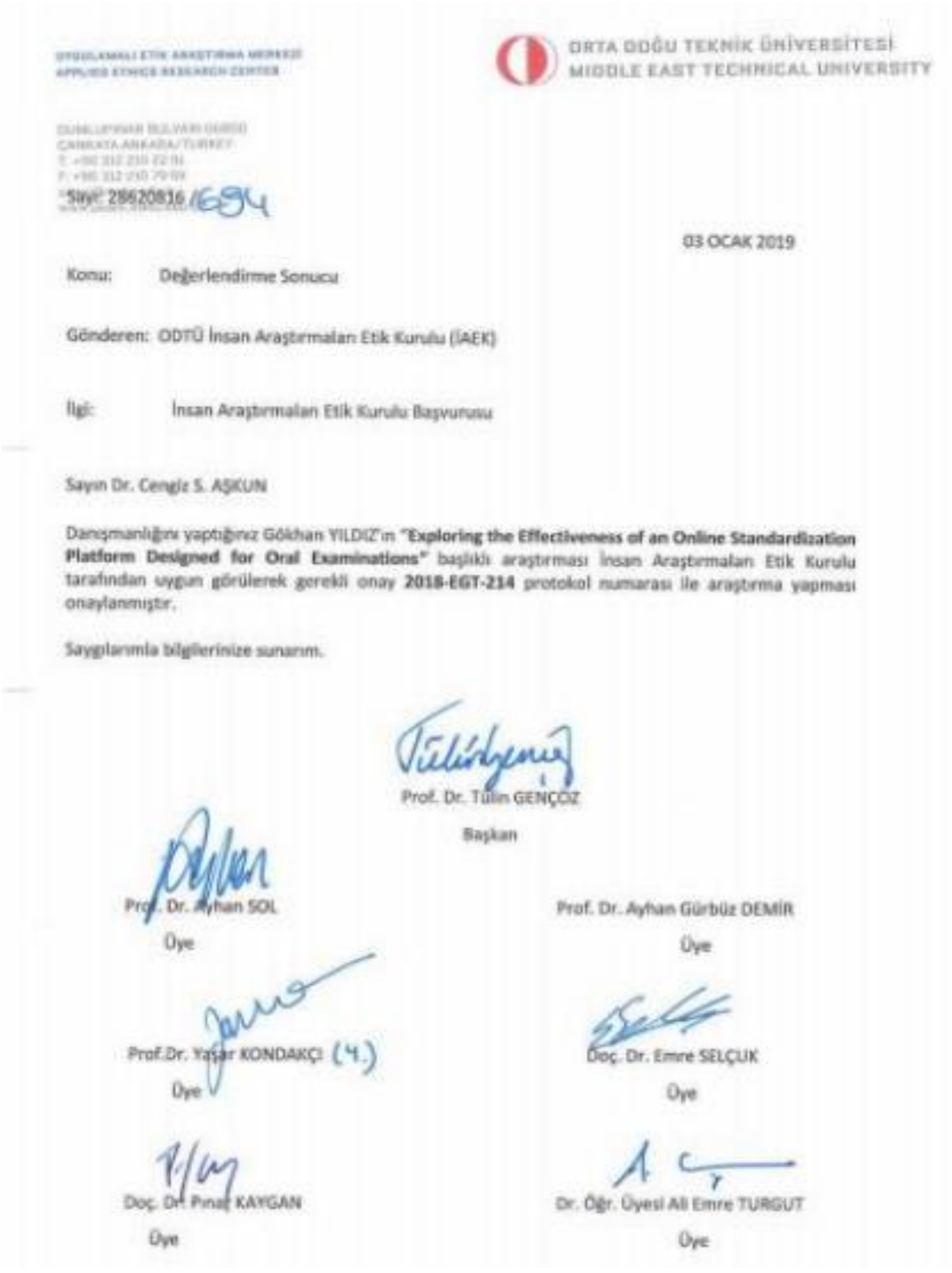
Zhao, Z. (2013). Diagnosing the english speaking ability of college students in China - Validation of the Diagnostic College English Speaking Test. *RELC Journal*.

APPENDICES

A. Semi-structured Interview Questions

- Aldığınız standardizasyon eğitimi hakkında ne düşünüyorsunuz?
- Aldığınız standardizasyon eğitiminde neleri iyi buldunuz?
- Aldığınız standardizasyon eğitiminde herhangi bir zorlukla karşılaştınız mı veya neleri yeterli bulmadınız? Nasıl değişiklikler yapılmasını isterdiniz?
- Aldığınız standardizasyon eğitimi sonrasında ilerde yapılacak olan konuşma sınavlarında değerlendirici olarak iyi bir performans göstereceğinizi düşünüyor musunuz?
    - Evet ise nedenini açıklar mısınız?
    - Hayır ise nedenini açıklar mısınız?
- Eklemek istediğiniz başka bir şey var mı?

B. Permission Obtained From METU Applied Ethics Research Center



UYGULAMALI ETİK ARAŞTIRMA MERKEZİ
APPLIED ETHICS RESEARCH CENTER

ORTA DOĞU TEKNİK ÜNİVERSİTESİ
MIDDLE EAST TECHNICAL UNIVERSITY

DUMLUPINAR BULVARI 06800
ÇANKAYA ANKARA/TURKEY
T: +90 312 210 22 91
F: +90 312 210 79 59

Sayı: 28620816 / 694

03 OCAK 2019

Konu: Değerlendirme Sonucu

Gönderen: ODTÜ İnsan Araştırmaları Etik Kurulu (İAEK)

İlgi: İnsan Araştırmaları Etik Kurulu Başvurusu

Sayın Dr. Cengiz S. AŞKUN

Danışmanlığını yaptığınız Gökhan YILDIZ'ın **"Exploring the Effectiveness of an Online Standardization Platform Designed for Oral Examinations"** başlıklı araştırması İnsan Araştırmaları Etik Kurulu tarafından uygun görülerek gerekli onay **2018-EGT-214** protokol numarası ile araştırma yapması onaylanmıştır.

Saygılarımla bilgilerinize sunarım.

Prof. Dr. Tülin GENÇÖZ
Başkan

Prof. Dr. Ayhan SOL
Üye

Prof. Dr. Ayhan Gürbüz DEMİR
Üye

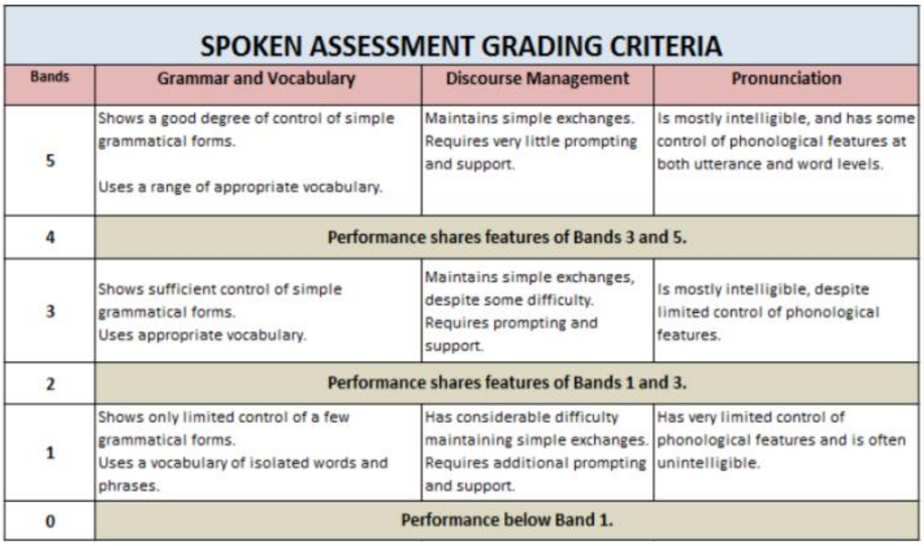
Prof.Dr. Yaşar KONDAKÇI (4.)
Üye

Doç. Dr. Emre SELÇUK
Üye

Doç. Dr. Pınar KAYGAN
Üye

Dr. Öğr. Üyesi Ali Emre TURGUT
Üye

C. Grading Criteria of the Oral Examination (Also used in standardization trainings)



SPOKEN ASSESSMENT GRADING CRITERIA

Bands	Grammar and Vocabulary	Discourse Management	Pronunciation
5	Shows a good degree of control of simple grammatical forms. Uses a range of appropriate vocabulary.	Maintains simple exchanges. Requires very little prompting and support.	Is mostly intelligible, and has some control of phonological features at both utterance and word levels.
4	Performance shares features of Bands 3 and 5.		
3	Shows sufficient control of simple grammatical forms. Uses appropriate vocabulary.	Maintains simple exchanges, despite some difficulty. Requires prompting and support.	Is mostly intelligible, despite limited control of phonological features.
2	Performance shares features of Bands 1 and 3.		
1	Shows only limited control of a few grammatical forms. Uses a vocabulary of isolated words and phrases.	Has considerable difficulty maintaining simple exchanges. Requires additional prompting and support.	Has very limited control of phonological features and is often unintelligible.
0	Performance below Band 1.		

Adapted from Cambridge University

D. Transcriptions of Cited Interviews

• Quotations used in 4.2.1 Positive Aspects about the Training Mediums Training Content

[FFS10] Örnek sınavları değerlendirmeden önce sınavla alakalı bilgiler almak gerçekten iyi oldu.

[OS1] İçerik olarak videoları ben gayet böyle "exam procedure" hakkında olsun ya da işte neyin ne sırayla yapılması gerektiğini, nasıl yapılması gerektiği konusunda çok başarılı buldum…Böyle kayboldum, devam edemeyeceğim şeklinde bir şeye de rastlamadım. O açıdan iyiydi… Herşey gayet açıklayıcı ve netti.

[OS4] Öncelikle bu 3'e bölünmüştü eğitim süreci. İlkinde yeteri kadar bilgi vardı bence. Videolar falan çok netti. Benim aklıma düşen bütün sorular vardı orada. Oradan bilgiyi alabildim.

[OS5] Bence içerikte herhangi bir sıkıntı yoktu herşey gayet anlaşılırdı ve takip etmesi kolaydı. Çok fazla detay ve ekstra bilgi yoktu.

[OS8] İçerik açısından bir sıkıntı yoktu ben eğlendim açıkçası. Bu standardizasyon sayesinde sınavın başlangıcından bitişine kadar ne yapacağımı öğrendim. Bu eğitim bütün gerekli bilgileri verdi.

[FFS1] Örnekler gayet kaliteliydi. Böyle ekstrem örnekler de gördük. Kendi öğrencilerimizden bekleyeceğimiz şeyleri gördük aslında ve insanların buna nasıl puan verdiklerini gördük. Bu baya işimizi kolaylaştıracak olan bir şeydi.

[FFS4] 10 tane örnek sınavın çeşitli seviyelerden seçilmesini beğendim. Hepsi aynı olsaydı manasız ve sıkıcı olacaktı bir yerden sonra ama farklı hataları farklı problemleri olan öğrencilerin seçilmesi hoşuma gitti.

[FFS5] Önce örnekler gördük, sample sınavları gördük nasıl yapıldığını. O iyiydi.

[OS1] Bence videoların hazırlanma şekli, sample exam'lerdeki o videoların konulması, çocuklar konuşurken picture description kısmının "embed" bir şekilde koyulması falan gayet iyiydi.

[FFS6] Bize burada geldiğimiz zaman, görsel bir şekilde kısımları mesela (warm-up kısmı) orda ne vardı onu gösteriyor. İkincinde resmi görüyoruz. Yani bir sınavdaymış gibi, yani o gerçekçiliği vermesi güzel çünkü geçtiğimiz senelerde de biz standardizasyon toplantıları olduğu zaman sadece ses kaydı dinlemiştik. O çok daha zordu. Burada ama öğrenci gerçekten karşımızdaymış gibi o deneyimi daha iyi yaşadık ve daha sağlıklı not verdik.

[FFS8] Sınavların bir video içerisinde verilmesi iyiydi. Öğrenci konuşurken öğrencinin neler hakkında veya hangi resim hakkında resmin kendisini videoda görmemiz oldukça iyiydi.

[FFS10] Öğrencilerin sınavlarda hakkında konuştukları soruların, fotoğrafların ve konuların videonun içinde görülebilmesi öğrenci konuşuyorken çok etkili oldu.

[OS4] Mesela videolar iyiydi. Açıklama yapan videolar. Çok uzun değildi. Çok ekstra bilgi yoktu içinde. Gereksiz bilgi yoktu. Hepsi nokta atışıydı diyebilirim.

[OS5] Videolar gerçekten çok iyi hazırlanmıştı.
[OS7] **...**ve sesli bir video değildi sadece yazıları okudum ve hani okuduğum şeyi kafamda tutmam daha da kolay olabiliyor bazen. O açıdan güzeldi.

Implementation of the Training

[OS1] Aldığım feedback'te burda böyleydi o yüzden böyle olması gerekiyor. Demek ki buna dikkat etmem gerekiyor şeklinde orada gelen feedback gayet açıklayıcı ve netti.

[OS9] Feedbacklerde böyle herşeyin tek tek açıklanması güzeldi bence. Örneklerle onu daha net görebildim bence.

[OS10] Örnek sınavlardan sonra değerlendirmelerin olduğu feedback sayfalarının gelmesi hani gayet iyi oldu çünkü öğrencilerin sınavlarında nelere dikkat etmek gerekiyor onları iyi bir şekilde görmüş oldum.

[OS11] Sample graded sınavların sonrasındaki feedbacklerde nasıl değerlendirmemiz gerektiği, nelere dikkat etmemiz gerektiğini görmemiz güzeldi.

[FFS2] Hani grupça beraber düşünmek ve bunun doğrultusunda feedback almak bana yardımcı oldu. Eğer böyle bir şey almasaydım büyük ihtimalle bocalardım.

[FFS3] Siz öncesinde bize bir bilgilendirme yaptınız sonrasında 3 tane örnek değerlendirilmiş sınavı yaptık. Daha önce verilmiş olan puanlar bizim nasıl değerlendirmemiz ya da cevaplarımızın o aralığa göre nereye olduğuna dair bir fikir verdi bize.Yani bir sınavı yaptıktan sonra beraber değerlendirdikten sonra diğer sınavı yapmamız ve tekrar beraber değerlendirmemiz güzel olabilirdi. Fakat bu da zaman açısından bize sıkıntı çıkarabilirdi.

[FFS4] Ama biraz daha uzun tartışabilseydik, herkesin başka işleri ve görevleri olduğu için çok yapamadık ama ben biraz daha tartışıp hani neden gerçekten böyle düşünülüyor vs. daha uzun tartışmak isterdim.

[FFS7] O şekilde standardizasyon yani feedback benim en çok hoşlandığım şeylerden birisi. Feedback olduğu için tabi ki daha etkili oldu.

[FFS9] Bir de tartışma ortamının olması çok iyiydi. Sonuçta standardize olabilmek için diğer hocaların da ne düşündüğüne ihtiyaç duyuyoruz. O 170 yüzden değerlendirmelerimizi konuşmamız daha önemli. Ama beraber değerlendirmediğimiz o sınavlardan sonrasında da beraber değerlendirme yapabilirdik.

[FFS2] Bence 2 oturum olması güzel çünkü bizim objektif olarak değerlendirmemizi artırıyor. Eğer bütün standardizasyon tek oturumda olsaydı bir yerden sonra kıyaslamaya başlıyoruz, motivasyonumuz düşüyor, değerlendirme kapasitesi de düşüyor. O yüzden 2 oturumda olması benim açımdan iyi oldu.

[FFS3] İki gün olduğu için bence çok iyi oldu motivasyon ve yorulmamak açısından. O yüzden bence idealdi.

[FFS6] Süresi bence uygundu biz iki tane oturum yaptık. İkisinde de çok aşırı, bizim çok zamanımızı da almayan aynı zamanda yeterli bir şekilde de (45-50 dk). Zaten böyle olması gerekiyordu. Aksi takdirde çok kısa olur veya çok uzun olursa da belki işin ana noktaları kaçabilir. Bence süre olarak da uygundu.

[FFS9] 2 oturuma yayılması kesinlikle daha sağlıklıydı. Sonuçta odaklanacağımız, dikkat edeceğimiz bir değerlendirme süreci söz konusu. Ve bu kadar dikkat edeceğimiz bir zamanda çok çok uzaması hocalar açısından birazcık sıkıntı olabilirdi.

[OS1] Süre açısından da bence ne çok uzun ne çok kısaydı. O açıdan iyiydi.

[OS2] İstediğimiz her zaman kolaylıkla erişebileceğimiz bir platformun olması rahatlatıcı.

Design of the Online Standardization Platform

[OS2] …kriterin solda videonun sağda olması çok iyiydi, pratikti.

[OS6] Şema olarak hem videoyu hem kriteri hem de puanlamayı aynı sayfada görmem güzeldi.

[OS8] İlk başta kriteri gördüm sonra ekran görüntüsünü aldım ve bir word dosyasına kopyaladım bir daha göremem ben bunu diye. Sonra sınavın yanında da olduğunu gördüm. O rahatlattı o yüzden. Öğrencinin sınavını dinlerken bir yandan da kritere bakabiliyor olmam çok iyiydi.

[OS10] Benim gözüme çarpan başka birşey de platformun olduğu sayfaların tam ekran gibi olması ve renklerin sade olması da gayet güzeldi.

[OS12] Yani genel olarak beğendim ben herşeyi. Kolay bir şekilde bilgi edinebildim. Çok fazla dolu olmadığı için site herhangi bir problem yaşamadım.

[OS3] Herşeyin aşama aşama gösterilmesi çok iyiydi. Sadece notları submit etmeden önce geriye dönebiliyoruz veya next diyerek tamamen submit edebiliyoruz. O güzeldi az önce değinmeyi unuttum.

[OS5] Web sitesi çok güzel hazırlanmış diye düşünüyorum. Mesela herşey adım adım gidiyor ve neyi nerede bulacağımız belli.

[OS9] İçerik de çok güzeldi çünkü 3 kısıma ayrılmıştı.

[OS10] Platformda 3 ana bölüm vardı. Bir menüye tıklayınca bilgilerin o menünün altına gelmesi gayet iyiydi. Sınav içeriği hakkında bilgi verilmesi ve eğitimin 3'e bölünmesi gayet iyiydi.

[OS1] Ve o sayfadaki işte puan verirken orada toplam puanı görüyor olabilmek, soruları görüyor olabilmek bunların hepsi bence önemliydi.

[OS2] … aşağıda verilen puanın direkt 100'lük sisteme çevirilmesi, puan hesap işleriyle uğraşılmaması çok iyiydi, pratikti.

[OS5] En çok puanlama kısmı hoşuma gitti. Kritere göre notları verince otomatik çevirmesi bence çok başarılı bir uygulama.

[OS10] … aşağıdaki kısımda notlandırma yapınca direkt olarak 100'lük sisteme göre öğrencinin kaç alıyor olduğunu görmek gayet iyiydi.

Practicality and Flexibility of the Platform

[OS2] Bir kere okula yeni başlayan hocalar için ve bizim için, yapılan prosedürü hatırlamamız için her zaman erişime açık bir yer olması. Zaman mekan sınırlamasının ortadan kaldırılması büyük bir avantaj.

[OS4] Ama sonrasında geri dönüp düşündüğümde elimde materyale anında erişebildiğim için bilgilere çok daha kolay oldu. Birine sormaktansa, bilgiyi aramaktansa elimin altında olduğu için sonrasında hiç problem yaşamadım mesela verdiğim puanlar benzer çıktı hep sample sınavlarla.

[OS5] Değerlendirmelerimi istediğim zaman yapabiliyor olmam bence en hoşuma giden kısım oldu.

OS6] … istediğimiz zaman mola veriyor olabilmemiz güzel.

[OS7] Bence çok güzel bir şey. Böyle bir şeyin online'a dönüştürüp zaman kazanmak çok faydalı ve mantıklı bir şey.

[OS3] Online olması açısından şöyle bir geriye dönebiliyoruz, not alırken durdurabiliyoruz. O anlamda çok daha fazla artısının olduğunu düşünüyorum. Ve çok daha uygulanabilir olduğunu düşünüyorum. Bir sonraki standardizasyonda tekrar online olarak değerlendirmeyi tercih ederdim.

[OS7] Çok güzel olduğunu düşünüyorum açıkcası. Online bir platformda olması güzeldi. Şöyle hani, kendi ev sürecimden biraz zaman harcamış oldum. O biraz başta canımı sıktı ama genel olarak hani sınava dair bir endişem kalmadı.

[OS12] Herşey çok güzeldi. Online şekilde olması da benim açımdan iyi oldu çünkü evdeyken yaptım hepsini. Okulda kalmak zorunda olmamam iyi bir şey diye düşünüyorum.

- **Quotations used in 4.2.2. Concerns, Weaknesses, and Suggestions to Improve the Standardization Mediums Concerns and Suggestions for the Online Standardization Platform**

[OS2] İlk başlarda puan farklılıkları yaşadım. Ama altında bu arada bir ekleme yapmak istiyorum. Katılmadığımız puanların admin'e yönlendirilmesini istiyorum sebebini yazarak. Hani bir itiraz butonu olmalı. Atıyorum bu puana 87'yi uygun görmüş sistem ama ben bunu 70 diye düşünüyorum. Şu sebeple düşünüyorum diye bir feedback kısmı olsaydı çok iyi olurdu.

[OS7] Sadece bu feedback kısmında şey oldu; sizin verdiğiniz puanlarla bazen benimkiler çakıştı. O zamanda benim danışabileceğim biri olmadı (in person).

Karşımda siz olsaydınız ben şey diyebilirdim: Hocam ben burada bunu verdim onu siz niye verdiniz diyerek direkt feedback alabilirdim.

[OS10] Platform genel olarak güzel hazırlanmış ama belki online bir şekilde feedback almadan yüz yüze olarak sizinle birkaç sınavı beraber değerlendirebilirdik. Direkt herşeyi websitesinden yaptığımız için bazen aklıma takılan şeyler oldu.

[FFS9] Sonuçta standardize olabilmek için diğer hocaların da ne düşündüğüne ihtiyaç duyuyoruz. O yüzden değerlendirmelerimizi konuşmamız daha önemli. Ama beraber değerlendirmediğimiz o sınavlardan sonrasında da beraber değerlendirme yapabilirdik.

[OS1] Mesela sample verdikten sonra bir yorum şeklinde bir hoca tarafından dinlemek belki güzel olabilirdi. Biz mesela önce feedback'i görüyoruz. Feedback'te böyle böyle olması gerekiyor.

[OS7] Belki sadece dediğim gibi sizinle görüşmek ve online'ı birleştirebilirdik hani. Sizinle biraz görüşebilirdik bir de online işlemi yapıp.

[OS10] … belki online bir şekilde feedback almadan yüz yüze olarak sizinle birkaç sınavı beraber değerlendirebilirdik.

Concerns and Suggestions for Face-to-face Standardization Platform

[FFS10] Zaman kısıtlamasından dolayı böyle olduğunu biliyorum ama daha geniş bir zaman aralığında belki yapılabilir bu.

[FFS3] Önceden değerlendirilmiş sınavdan 3 tane yaptık. 10 tane de sonrasında yaptık. Bence gayet iyiydi. Siz öncesinde bize bir bilgilendirme yaptınız sonrasında 3 tane örnek değerlendirilmiş sınavı yaptık. Daha önce verilmiş olan puanlar bizim nasıl değerlendirmemiz ya da cevaplarımızın o aralığa göre nereye olduğuna dair bir fikir verdi bize. Bence eksik bir şey yoktu. Ama sadece en son kendimizin değerlendirdiği 10 örnek sınavı beraber değerlendirmemiz güzel olabilirdi.

[FFS4] … Ama biraz daha uzun tartışabilseydik, herkesin başka işleri ve görevleri olduğu için çok yapamadık ama ben biraz daha tartışıp hani neden gerçekten böyle düşünülüyor vs. daha uzun tartışmak isterdim.

[FFS9] … Ama beraber değerlendirmediğimiz o sınavlardan sonrasında da beraber değerlendirme yapabilirdik.