THE REPUBLIC OF TÜRKİYE
SAKARYA UNIVERSITY
INSTITUTE OF EDUCATIONAL SCIENCES
ENGLISH LANGUAGE TEACHING DEPARTMENT

VALIDITY AND RELIABILITY OF EFL SPEAKING TESTS IN TURKISH STATE HIGH SCHOOLS

MASTER'S THESIS

MUHAMMED AK

SUPERVISOR
ASSIST. PROF. DR. ELİF BOZYİĞİT

OCTOBER 2025

THE REPUBLIC OF TÜRKİYE
SAKARYA UNIVERSITY
INSTITUTE OF EDUCATIONAL SCIENCES
ENGLISH LANGUAGE TEACHING DEPARTMENT

VALIDITY AND RELIABILITY OF EFL SPEAKING TESTS IN TURKISH STATE HIGH SCHOOLS

MASTER'S THESIS

MUHAMMED AK

SUPERVISOR

ASSIST. PROF. DR. ELİF BOZYİĞİT

OCTOBER 2025

DECLARATION

I declare in this study, which I prepared in accordance with the Sakarya University Institute of Educational Sciences Thesis Writing Guide, that:

- All information and documents included in the thesis have been obtained and presented in accordance with academic and ethical principles,
- I have cited and referenced all sources I have benefited from,
- I have not made any alterations to the data used,
- I have not submitted this thesis, in whole or in part, as another thesis study.

Muhammed AK

JÜRİ ÜYELERİNİN İMZA SAYFASI

Muhammed AK tarafından hazırlanan **"Validity and Reliability of EFL Speaking Tests in Turkish State High Schools"** adlı tez çalışması 10/10/2025 tarihinde aşağıdaki jüri tarafından oy birliği ile Sakarya Üniversitesi Eğitim Bilimleri Enstitüsü **Yabancı Diller Eğitimi Anabilim Dalı İngiliz Dili Eğitimi Bilim Dalı'nda Yüksek Lisans tezi** olarak kabul edilmiştir.

Tez Jürisi

Jüri Başkanı:	Dr. Öğr. Üyesi CEMİL GÖKHAN KARACAN Beykent Üniversitesi
Jüri Üyesi (Danışman):	Dr. Öğr. Üyesi ELİF BOZYİĞİT Sakarya Üniversitesi
Jüri Üyesi:	Dr. Öğr. Üyesi ALİ İLYA Sakarya Üniversitesi

ACKNOWLEDGEMENT

As an actively working teacher with numerous responsibilities, balancing the demands of professional and academic life has been a challenging journey. This path, though demanding, has helped me grow both personally and professionally. I would like to take this opportunity to thank a few who have supported me throughout this journey.

I owe my deepest thanks to my dear wife, Şule, whose love and patience carried me through every step of this process.

To my little daughter, Elif Rana. Your innocent questions, hugs, and jealousy of the time I spent on my studies always reminded me of the balance I should preserve between my work, academic life and family.

Finally, I am thankful to my supervisor, Assist. Prof. Dr. Elif BOZYİĞİT, my classmate Harun KARAŞ, all participating teachers and contributing academicians for their valuable input in this study.

ABSTRACT

VALIDITY AND RELIABILITY OF EFL SPEAKING TESTS IN TURKISH STATE HIGH SCHOOLS

Muhammed AK, Master's Thesis

Supervisor: Assist. Prof. Dr. Elif BOZYİĞİT

Sakarya University, 2025.

In foreign language education, classroom-based summative assessment practices are a fundamental part of the process that guides instruction and allows monitoring of student development. Since there is currently no centralized four-skill English proficiency test in Türkiye, what is known about students' speaking ability largely depends on the assessments designed and administered by teachers in schools. According to the 2023 Directive for Written and Practical Exams published by the Ministry of National Education, a major change was introduced in the national assessment system. English as a foreign language tests are now conducted in two parts: written and practical. While the written test focuses on reading and writing skills, the latter focuses on speaking and listening skills. Given that teacher-led assessments may pose challenges in ensuring validity and reliability, this study analyzes the validity and reliability properties of 12 high school English speaking tests across four common school types in Türkiye. By focusing on key aspects such as face validity, content validity, construct validity, and reliability, the study seeks to reveal strengths and weaknesses in current classroom-based speaking tests. Data were collected through expert evaluation using a checklist and analyzed through statistical methods. The findings indicated that while the speaking tests exhibited satisfactory levels of face and content validity, their construct validity remained notably limited. Overall reliability was also limited, largely due to the absence of detailed scoring rubrics, inconsistent criteria, and a lack of moderation procedures. Findings suggest that classroom-based speaking tests in Turkish secondary schools are shaped by practical constraints rather than standardized assessment frameworks. The outcomes underscore the need for structured professional development in speaking assessment, the development of analytic rubrics aligned with curricular standards, and the establishment of institutional mechanisms to ensure fairness and consistency across school contexts.

Anahtar Kelimeler: Testing, Speaking, Reliability, Validity, English Language Teaching.

ÖZET

TÜRKİYE'DEKİ DEVLET LİSELERİNDE İNGİLİZCE KONUŞMA TESTLERİNİN GEÇERLİĞİ VE GÜVENİRLİĞİ

Muhammed AK, Yüksek Lisans Tezi

Danışman: Dr. Elif BOZYİĞİT

Sakarya Üniversitesi, 2025.

Yabancı dil eğitiminde sınıf içi özetleyici ölçme-değerlendirme uygulamaları, öğretimi yönlendiren ve öğrenci gelişimini izlemeye olanak tanıyan sürecin temel bir parçasıdır. Türkiye'de şu anda merkezi bir dört beceriye dayalı İngilizce yeterlik sınavı bulunmadığından, öğrencilerin konuşma becerilerine ilişkin bildiklerimiz büyük ölçüde okullarda öğretmenler tarafından tasarlanıp uygulanan değerlendirmelere dayanmaktadır. Milli Eğitim Bakanlığının 2023 tarihli Yazılı ve Uygulamalı Sınavlar Yönergesi ile ölçme-değerlendirme sisteminde önemli bir değişiklik yapılmış; İngilizce sınavları artık yazılı (okuma-yazma) ve uygulamalı (konuşma-dinleme) olmak üzere iki bölüm hâlinde uygulanmaya başlanmıştır. Öğretmen tarafından yürütülen bu uygulamalı sınavların geçerlik ve güvenirlik açısından çeşitli sorunlar barındırabileceği değerlendirilmektedir. Bu çalışma, Türkiye'de yaygın olarak bulunan dört farklı yaygın lise türünde uygulanan 12 İngilizce konuşma sınavının psikometrik özelliklerini incelemektedir. Araştırma, görünüş geçerliği, kapsam geçerliği, yapı geçerliği ve güvenirlik gibi temel ölçme değerlendirme boyutlarına odaklanarak mevcut konuşma becerisi testlerinin güçlü ve zayıf yönlerini ortaya koymayı amaçlamaktadır. Veriler bir kontrol listesi aracılığıyla uzman değerlendirmeleriyle toplanmış pek çok istatistik yöntemleri ile çözümlenmiştir. Bulgular, konuşma sınavlarının görünüş ve kapsam geçerliği açısından tatmin edici düzeyler sergilediğini, ancak yapı geçerliğinin belirgin biçimde sınırlı kaldığını göstermiştir. Genel güvenirlik düzeyi de sınırlı bulunmuş; bu durum, ayrıntılı puanlama rubriklerinin olmaması, ölçütlerdeki tutarsızlıklar ve değerlendirme sürecinde moderasyon uygulamalarının eksikliğinden kaynaklanmıştır. Bulgular, Türkiye'deki liselerde konuşma sınavlarının, standartlaştırılmış ölçme-değerlendirme çerçevelerinden ziyade pratik kısıtlamalar tarafından şekillendirildiğini ortaya koymaktadır. Elde edilen sonuçlar, konuşma becerisi değerlendirmesi alanında yapılandırılmış mesleki gelişim programlarının gerekliliğini, öğretim programı standartlarıyla uyumlu analitik rubriklerin geliştirilmesini ve okul bağlamlarında adalet ve tutarlılığı sağlamak üzere kurumsal mekanizmaların oluşturulması ihtiyacını vurgulamaktadır.

Keywords: Ölçme, Konuşma Becerisi, Güvenirlik, Geçerlik, İngiliz Dili Eğitimi.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

CEFR: Common European Framework of Reference for Languages

TOEFL: Test of English as a Foreign Language

MoNE: Ministry of National Education

IBT: Internet-Based Test

ETS: Educational Testing Service

MA: Master of Arts

EFL: English as a Foreign Language

ESL: English as a Second Language

IC: Interactive Competence

TEPSE: Transition Examination from Primary to Secondary Education

ELT: English Language Teaching

BERA: British Educational Research Association

CVI: Content Validity Index

ANOVA: Analysis of Variance

SPSS: Statistical Package for the Social Sciences

SD: Standard Deviation

CHAPTER I

INTRODUCTION

English language has reached a crucial value due to the increased international communication and globalization (Friedman, 2005; Majidifard, Shomoossi and Ghourchaei, 2014), international trades and career opportunities (Clement and Murugavel 2018; Crystal, 2003; Graddol, 2006) and the need for cultural exchange and understanding (Durai and Soundrarajan, 2009; Dutta, 2019; Hubackova, 2016; Rao, 2019).

Teaching English as a foreign language in Türkiye gained momentum in the mid-twentieth century and became increasingly popular with the intensive language teaching programs and English-medium instruction in some high schools, which was visible in Education First's Index results (Rajan, 2013). Although its intensity and the educational levels at which it is taught have changed over time due to various alterations, it has largely maintained its importance in terms of both instructional time and value within the education system.

In response to the importance placed on the subject, many studies conducted in the first decade of the 2000s focused on the failure to achieve communicative outcomes in language teaching in Türkiye, and tried to offer solutions to this issue (Gökdemir, 2010; Işık, 2005, 2008; Kızıldag, 2009; Paker, 2007; Tılfarlıoglu and Öztürk, 2007). Since the second decade of the 2000s, the transition to the Common European Framework of Reference for Languages (CEFR) as part of the standardization of the language teaching process has become more visible in the field and skill-based approaches have begun to be implemented both in universities and in schools under the Ministry of National Education (MoNE). Today, the learning outcomes of courses are expressed separately as reading, writing, listening, and speaking, and all four skills are included in the curriculum and reflected in the textbooks used in public schools.

1.1.Problem statement

Speaking can be defined as the oral use of language for communication with others (Fulcher, 2003). Despite this central role of speaking in communication, learners often struggle to

develop it adequately (Bailey, 2005; Luoma, 2004; Nunan, 1999). When it comes to English, the phrase 'I understand, but I can't speak' is often heard in Türkiye. This statement, which highlights the lack of speaking skill instruction, is also reflected in the statistics of highly valid international exams. According to the Test of English as a Foreign Language (TOEFL) Internet-Based Test (IBT) Test and Score Data Summary (2023) published by the Educational Testing Service (ETS), it is observed that students whose native language is Turkish receive the lowest scores in the speaking section among the four language skills.

Speaking skills offer some peculiar advantages over other language skills. It plays a motivating role in language learning as it reinforces the interactions in the target language, enabling increased use of the language meaningfully (Aliqi and Hayes, 2024). Also, students with good speaking skills are known to have better social relationships and find more opportunities in their careers (Akhter, 2021). Besides, tone of voice and intonation can only be used meaningfully while speaking. Moreover, one can receive immediate feedback from the listener while expressing his/her ideas, feelings, etc. directly through speaking.

As assessment reflects both student progress and teaching effectiveness, accurate evaluation of speaking skills is vital. In recent years, the evaluation of speaking ability has gained prominence as one of the major issues in language testing (Sak, 2008; Morrow, 2012). Although language assessment has drawn considerable interest from both practitioners and researchers, the time and emphasis devoted to it are still insufficient (Hatipoğlu, 2015).

Data obtained through assessment and evaluation practices play a crucial role in enhancing and advancing all dimensions of the educational process (Doğru, 2019, p.1). Moreover, the assessment and evaluation process is essential for determining students' current achievement levels, identifying areas within the education system that are in need of improvement, detecting deficiencies, and guiding necessary developmental initiatives (Akıncı, 2020). Wigglesworth (1993) emphasizes the need for the test designers to focus on the language outcomes and assess the language skills to the fullest.

Assessment and evaluation provide an opportunity to review, revise, and even entirely reform or eliminate processes within the education system itself (Baykul, 2010, p. 97), which requires assessment to be a meticulous process. In order to conduct proper assessment practices, teachers are expected to be competent in this field (Herrera and Macias, 2015). However, studies indicate inefficiencies regarding the English language teachers' assessment skills in Türkiye (Ballıdağ and İnan Karagül, 2021; Genç, Çalışkan and Yüksel,

2020; Önalan, 2020; Semiz and Odabaş, 2016; Valizadeh, 2019). Several studies have indicated that the 'Assessment and Evaluation in Foreign Language Teaching' course offered at universities does not provide teacher candidates with sufficient training in assessing language skills (Hatipoğlu, 2015; Haznedar, 2010; Öz, 2014). Besides, studies indicate that teacher candidates also have limited opportunities to apply theoretical knowledge in practice (Hasselgreen, Carlsen and Helness, 2004; Köksal, 2004; Karakaş, 2012; Vogt and Tsagari, 2014). Brown and Baile (2008) and Coombe, Troudi, and Al-Hamly (2012) reported that teachers, due to insufficient training, experience anxiety about assessment, view it as separate from teaching, and consequently feel inadequate. This lack of training was visible in İleri's study (2019), where she stated that between 2008 and 2018, no in-service training activities were organized to improve English teachers' assessment and evaluation skills. However, it has been observed that, as of 2021, new courses have been introduced to help English teachers improve their assessment competencies (MEB, 2024).

Similarly, it has been found that there are several assessment practices contradicting the English Language Teaching Program of MoNE (Akçay, Tunagür and Karabulut, 2020; Arslan and Üçok-Atasoy, 2020; Çimen, 2022). According to McDonald (2002) and Airasian and Russel (2012), sound decision-making in educational contexts and the accurate measurement of academic achievement can only be achieved through reliable and valid assessments. The results of above mentioned studies raise concerns about reliability and validity in exams.

Reliability refers to the extent to which the results obtained from a measurement instrument are free from random errors (Turgut, 1990). A reliable exam will produce similar results when repeated under similar circumstances. Reliable exams reduce the impact of random errors such as temporary conditions of students (stress, tiredness, etc.), factors related to the setting (noise, temperature, etc.) and the subjectivity of the rater (previous thoughts about students, biases, etc.) (Brown 1996, p. 198). Moreover, educators and other stakeholders, such as policymakers, require reliable results to make certain decisions.

Validity refers to the accuracy of an assessment tool regarding the item it is intended to assess, without confusing it with any other characteristic (Tekin, 1977). Validity ensures that the conclusions drawn from the assessment results are meaningful, appropriate, and valid. An exam with poor validity may not reflect the reality of students' performance. For instance, a student may perform badly in a speaking test just because of the inappropriateness of the topic or criteria. Validity is also a must for fair grading and access to upper educational

institutions. Without a valid exam, teachers or educational policymakers cannot evaluate the effectiveness of the teaching. Finally, a valid assessment system helps teachers, school managers and policy makers in the accountability process.

In 2016, new regulations were established regarding the assessment of the four skills in foreign language education in Türkiye (Regulation for Secondary Education Institutions, 2016). Even so, studies showed multiple choice questions being used for assessing all skills, sentence completion questions being used for speaking assessment, no consensus among teachers on rubrics and indifference of speaking tests to classroom practices (Çimen, 2022; Özdemir, 2018).

In 2023, the MoNE published the Directive for Written and Practical Exams, making a major change in the assessment system. Since then, the English as a foreign language tests have been conducted in two separate versions as written and practical. The written part consists of reading and writing, while the latter focuses on listening and speaking skills.

To the best of the researcher's knowledge, although there are studies focusing on the assessment literacy of English teachers (Ballıdağ and İnan Karagül, 2021; Genç, Çalışkan and Yüksel, 2020; İleri, 2019), perspectives of students' and teachers' on speaking assessment (Özdemir, 2018), and the alignment of speaking tests with the curriculum (Çimen, 2022), there has not been any research focusing on the reliability and validity of the speaking tests conducted in high schools in Türkiye.

1.2.The purpose and importance of the study

This study aims to evaluate the speaking tests used by English language teachers in high schools in Sakarya in terms of validity and reliability.

This study may play a role in shaping educational policy by providing insights (regarding the validity and reliability aspects of EFL speaking tests) from the field to guide national standards, guidelines and reforms in foreign language assessment. It may also support curriculum development by comparing the curricular goals with the assessment practices. Moreover, the findings could contribute to institutional accountability, encouraging schools or the MoNE to take action towards a more equitable assessment system that promotes fairness for all learners.

This study provides evidence regarding the validity and reliability of teachers' current practices in language assessment, which may help improve teacher assessment literacy. The findings may affect pre-service teacher education programs by integrating the findings into teaching practices on language assessment. Additionally, it can also inform in-service teacher training and professional development courses.

Given the limited research on validity and reliability properties of foreign language speaking tests in Türkiye, this study provides specific empirical data on how speaking skills are assessed, opens up new research opportunities in the academic literature, and may help promote fairness and equity in language assessment in the long run.

1.3.Research Question

To what extent do the speaking tests conducted in high schools in Sakarya province demonstrate psychometric soundness in terms of validity and reliability?

1.4.Sub-questions

1. To what extent do the speaking tests conducted in high schools in Sakarya province demonstrate face validity?
2. To what extent do the speaking tests conducted in high schools in Sakarya province demonstrate content validity?
3. To what extent do the speaking tests conducted in high schools in Sakarya province demonstrate construct validity?
4. To what extent do the speaking tests conducted in high schools in Sakarya province demonstrate reliability?

1.5.Assumptions

It is assumed that the teachers participating in the study provided honest and accurate responses regarding the speaking tests. It is also assumed that the practices observed in the secondary education institutions included in the study represent the general situation. Furthermore, it is assumed that the experts participating in the evaluation possessed sufficient expertise and maintained objectivity throughout the process.

1.6.Limitations

The study is limited to four school types, each of which is represented by three different schools, totalling 12 high schools in Sakarya province. The research only focused on the second speaking tests for the tenth-grade in the first semester. Observation or video recording during the tests was not conducted by the researcher due to practical and economical reasons, so the research is limited to the information provided by the teachers regarding the test. However, the data collection instruments were carefully designed to minimize the impact of this limitation. The study does not provide evidence regarding response and washback validity, as no data were collected on how test-takers interpreted or approached the test tasks.

1.7.Definitions

Psychometrics: Psychometrics is a field that focuses on designing measurement instruments, testing their validity, and examining whether these instruments are appropriate in terms of reliability and validity (Ginty, 2020).

Speaking Test: Speaking tests are summative assessments prepared and conducted by teachers as a part of practical exams that are required by the Turkish National Ministry of Education.

Validity: The degree to which a test measures what it claims to be measuring (Brown, 1996).

Reliability: The extent to which the results obtained from a measurement instrument are free from random errors (Turgut, 1990).

Assessment literacy: A teacher's assessment competence (Stiggings, 1991).

CHAPTER II

THEORETICAL FRAMEWORK OF THE RESEARCH AND LITERATURE REVIEW

2.1. Language Skills and Assessment

Language skills are generally divided into two categories: receptive skills (reading and listening) and productive skills (speaking and writing) (Sreena and Ilankumaran, 2018). While receptive skills involve understanding input, productive skills require the learner to produce output. Each language skill's peculiar nature requires the development of specialized assessment tools with regard to that skill's characteristics.

2.1.1. Reading skills and assessment

Decoding and understanding letters may sound pretty basic for this century, but this was not the case for centuries. Today, worldwide literacy rates reach up to 90 per cent, and in many countries, it is above 99 per cent (UNESCO, 2024). Teaching reading is one of the fundamentals in education, which requires the development of specific assessment tools for its evaluation. Being one of the receptive skills, reading skills are not observed during practice, which requires all assessment of reading to be carried out by inference (Brown and Abeywickrama, 2010). Its assessment can be done relatively quickly and easily compared to other skills with pen-and-paper exams. Assessment of reading can focus on several sub-skills, from skimming and scanning to distinguishing fact from opinion. It can be assessed using many techniques such as multiple choice, short answer, gap filling and picture-cued tasks. However, one challenge in reading assessment is that tasks sometimes measure background knowledge rather than actual reading ability, which raises concerns about validity (Alderson, 2000). For this reason, careful task design is necessary to ensure that scores reflect reading skills rather than content familiarity.

2.1.2. Listening skills and assessment

Listening skill shows many similarities with reading, as it is also a receptive skill. The main difference may be the fact that while reading, one can go back and read again, but while listening, one cannot simply stop and listen again unless there is a recording. Another difference is that listening skills are usually accompanied by speaking skills. In most everyday situations, speaking and listening skills are integrated; however majority of the tests tend to focus on listening skills separately by using videos or audio tapes. Listening involves informational forms like following instructions and obtaining factual information, and interactional forms such as recognising requests for clarification and understanding expressions of disagreement (Hughes, 2003). Listening assessment also faces validity concerns, as tasks may unintentionally test short-term memory or familiarity with accents and cultural references rather than comprehension itself (Buck, 2001). These issues make task selection and input design crucial for fairness.

2.1.3. Writing skills and assessment

Being one of the productive skills, writing is one of the skills that you can observe the product. Although it requires a longer time to assess, it has been commonly assessed due to its convenience in pen-and-paper exams. Writing is usually assessed separately; however, reading may be integrated through long instructions such as scenarios. Brown and Abeywickrama (2010) explain writing in three genres: academic, job-related and personal writing, and in four categories: imitative, intensive, responsive and extensive. Each of these divisions requires unique assessment tools like dictation, paragraph completion, short answer and paragraph construction. Unlike reading or listening, writing often requires subjective scoring, which can reduce reliability unless rubrics and multiple raters are used (Weigle, 2002). Practicality is another concern, as evaluating writing takes considerably more time than receptive skills.

2.1.4. Speaking skills and assessment

The development of language testing has closely followed changes in how language and learning have been understood. Early examinations in the late 19th and early 20th centuries were shaped by the grammar-translation tradition, where the focus was on written translation and grammatical knowledge rather than on actual use of the language (Spolsky, 1995). By the mid-20th century, under the influence of structural linguistics and behaviorism, tests started to be in discrete-point formats that measured isolated elements such as grammar or vocabulary through highly controlled items (Lado, 1961). Still, as Weir (2003) notes, there was no clear model for assessing speaking before the mid-1970s, and the tasks often fell short of capturing oral ability. Dictation, for example, was more a measure of listening and writing, while reading aloud assessed pronunciation alongside low-level reading skills (O'Sullivan, 2013).

A major turning point came in 1975 with the revision of the Cambridge Proficiency Exam (CPE) oral paper. This change coincides with the rise of Communicative Language Teaching (CLT) and marks the beginning of more authentic approaches to speaking assessment. The emphasis shifted from testing language knowledge in isolation to evaluating how learners could use language for real communicative purposes (Canale and Swain, 1980). From the 1990s onwards, this orientation led to a stronger focus on performance-based testing, supported by frameworks such as the Common European Framework of Reference (CEFR) and the broader concept of communicative competence (Bachman, 1990; Fulcher and Davidson, 2007). Such developments in testing philosophy reflect the distinctive nature of speaking itself, which requires both linguistic control and the ability to respond in real time.

Speaking is a complex language skill that requires the ability to organize thoughts meaningfully while drawing on multiple linguistic elements such as pronunciation, intonation, vocabulary, and grammar. Its major distinction from other skills is its dynamic, interactive nature. Unlike writing, this productive skill usually takes place in real time, demanding quick thinking, responsiveness to interlocutors, and sensitivity to intonation and body language. For native speakers, oral language develops naturally from an early age and typically matures within a few years through constant exposure. In contrast, most non-native speakers begin learning a foreign language in primary school under low-exposure conditions. This disadvantaged context makes the teaching of speaking not only more challenging but also more crucial. These features highlight why the assessment of oral language deserves particular attention.

2.1.5. The Importance of Oral Language Testing

Language tests are widely recognized as essential instruments for evaluating learners' abilities in specific domains. Brown (2004) notes that they serve as tools for providing an accurate measurement of what students can do with the language. Beyond measurement, tests also play a formative role. Sheng-ping and Chong-ning (2004) describe them as a central source of feedback for both teachers and students, making it possible to reflect on teaching and learning practices in order to improve them. In this sense, tests are not simply mechanisms for assigning grades; they are a means of supporting the overall learning process.

Within the field of language testing, speaking skills have long been identified as one of the most important yet most complex skills to evaluate, which will be covered in the following chapter. Heaton (2003) underlines that testing the ability to speak is crucial but remains a very difficult task because of its multifaceted nature. The interactive and spontaneous qualities of speaking make it challenging to capture in a fair and reliable way, but its importance cannot be ignored. Also as Çaykan (2001) as cited in Özdemir (2018), observed, students tend to give more importance to oral skills when they expect an oral test, while they often neglect these skills when no such assessment is anticipated. Paker (2013) also reinforces this point by noting that learners are unlikely to overlook any skill as long as it is subject to evaluation. These findings highlight the role of speaking assessment in shaping learners' priorities and study habits.

The benefits of language testing extend beyond motivation. Madsen (1983) emphasizes that tests confirm progress for students, teachers, and administrators, helping all parties see the outcomes of teaching and learning. For students, tests provide a sense of accomplishment and reassurance that the teacher's evaluation corresponds with what has been taught. This connection between teaching and testing can foster positive attitudes toward instruction and create greater trust in the educational process. At the same time, tests highlight areas for improvement, giving learners direction for further study and allowing them to focus on their specific needs.

From the teacher's perspective, speaking assessment is equally valuable. Bachman and Palmer (1996) point out that test results inform a wide range of instructional decisions, including the choice of learning materials, the design of classroom activities, and the

diagnosis of student strengths and weaknesses. Teachers can use assessments to determine whether students are ready to advance to new content or whether they require additional support. Assessment also provides the basis for assigning grades in a way that is aligned with student achievement. In this way, speaking tests serve both diagnostic and summative purposes, guiding instruction while also evaluating learning outcomes.

The importance of speaking assessment also extends beyond the classroom. Douglas (2014) emphasizes that language tests provide assurance to various stakeholders, such as administrators, parents, admissions officers, and employers. Through tests, these groups gain confidence that learners are meeting recognized standards or reaching required levels of language proficiency. McNamara and Shohamy (2008) add that the influence of testing reaches into political and social contexts, reflecting the wider impact of assessment beyond individual classrooms. Tests therefore carry both educational and societal weight, reinforcing their significance.

Finally, speaking assessment plays a strong role in shaping teaching practices. As McEwen (1995) observes, what is tested tends to become what is valued, and what is valued becomes what is taught. This means that when speaking skills are assessed, they naturally gain importance within the curriculum and classroom practice. Conversely, if speaking is overlooked in assessment, it risks being neglected in instruction. This principle highlights the powerful role of speaking assessment not only in measuring learning but also in directing it.

Taken together, the literature suggests that speaking assessment is vital for learners, teachers, and broader educational stakeholders. It motivates students to develop oral skills, provides direction for learning and teaching, confirms progress, and assures external audiences of learners' competence. At the same time, it influences the curriculum by determining which skills are emphasized in instruction. Although assessing speaking is challenging, its importance makes it a necessary focus for both researchers and practitioners. This naturally leads to a discussion of the challenges associated with speaking assessment, which will be addressed in the following section.

2.1.6. Challenging Nature of Oral Language Testing

Madsen (1983) describes speaking assessment as the most challenging of all language tests to prepare, administer, and score, a view widely repeated in later research. Unlike written work, which produces a permanent record that can be re-examined, oral performance is ephemeral and disappears as soon as it is produced. This transience makes speaking unique. Roca-Varela and Palacios (2013) emphasize speaking is particularly demanding to evaluate because it requires extended time and contains features that are difficult to capture. Weir (2005) also points out that since many speaking tests are conducted without recording, raters must make immediate judgments, which increases the risk of inconsistency.

Hughes (2003) observes that language tests are often designed around what is easy to measure rather than what is pedagogically important. This tendency can distort assessment priorities, though careful content specifications may help mitigate the issue. Also, as Höl (2018) notes, the dual role of teachers as both assessors and interlocutors demands substantial time, effort, and impartiality, making it difficult to achieve reliable results.

The complexity of speaking assessment has been confirmed by multiple scholars. Fitzpatrick and Morrison (1971) argue that oral exams demand greater performance than traditional paper-and-pencil tasks. Luoma (2004) explains that speaking assessment is difficult because impressions of oral ability are shaped by numerous factors, yet scores are still expected to be accurate. Bachman (2004) further emphasizes that speaking scores are influenced not only by the learners themselves but also by the test tasks, raters, scoring criteria, and testing conditions.

At the same time, the act of rating itself is cognitively demanding. Raters must observe, interpret, recall, and integrate information in real time, processes that impose a heavy mental workload (Eckes, 2012; Myford and Wolfe, 2004; Tavares and Eva, 2013). Test takers also face challenges, since speaking requires them to plan, process, and produce language on the spot, often under pressure (Luoma, 2004). The type of interaction, the design of tasks, and the opportunities provided to demonstrate ability all significantly affect performance (Luoma, 2004).

Practical concerns further complicate speaking assessment. Sak (2008) draws attention to issues such as administrative costs, the difficulty of testing large groups individually or in pairs, and the training required for teachers. Similarly, numerous factors, including topic selection, test format, interlocutor behavior, and test-taker characteristics, can create

variability in scores (Brown, 2003, 2005; Berry, 2007; Brindley, 1991; O'Sullivan, 2006; Shohamy, 1988, 1994).

Another layer of complexity arises from the relationship between speaking and listening. Kitao and Kitao (1998) highlight that speaking cannot be separated from listening, since effective oral communication depends in part on comprehending spoken input. This makes it necessary for many assessments to integrate listening or even reading skills. While such integrated tasks increase authenticity, they also pose design and scoring challenges, and learners tend to perform less successfully on them than on discrete speaking tasks (Brown et al., 2001; Douglas, 1997; Weir, 1990). Because of these difficulties, both teachers and students sometimes avoid focusing on speaking, leaving it under-assessed compared to other skills (Allan, 1999; O'Malley and Valdez-Pierce, 1996).

Altogether, the literature consistently portrays speaking as the most complex area of language assessment. Its transient nature, the cognitive and practical demands on both raters and learners, and the logistical barriers involved all contribute to its difficulty. At the same time, these challenges point to the need for careful test design, robust rating criteria, and adequate training to ensure that speaking can be assessed fairly and meaningfully.

2.2. Psychometric Properties in Speaking Tests

Psychometrics is the construction and validation of measurement instruments and assessing if these instruments are reliable and valid forms of measurement (Ginty, 2020). In educational assessment, psychometric properties are of crucial importance in conducting effective and accountable assessment (Ohiri and Ihebom, 2024; Xobin, 2024). In this research, it is sought to portray the psychometric properties of speaking tests in terms of validity and reliability.

This study focused on face validity, content validity, construct validity, and reliability dimensions. By analyzing expert judgment statistically, the research aims to portray the current state of English as a foreign language speaking tests in state high schools in Sakarya province.

2.2.1. Validity in Speaking Tests

Validity can be defined as the degree to which a measurement tool can accurately measure the characteristic it aims to measure, without confusing it with any other variable (Tekin, 1977), the degree to which a test measures what it claims to be measuring (Brown, 1996), or the degree of agreement between a test score or measurement and the quality it is assumed to represent (Kaplan and Saccuzzo, 2001) or similarly, the extent to which inference made from the assessment results are appropriate, meaningful, and useful in terms of the purpose of the assessment (Gronlund, 1998 p.226 as cited in Brown and Abeywickrama, 2010).

Validity is widely regarded as the most important consideration in test development, since without validity, other qualities such as reliability lose their significance (Clark, 1978; Luoma, 2004). Carey (1988) emphasizes that a test can only be considered valid if it is designed in accordance with the intended learning outcomes and contains sufficient and appropriate items that accurately represent those outcomes. Kitao and Kitao (1998) explain that a test designed to measure communicative ability in English can only be considered valid if it truly assesses communication rather than grammatical knowledge.

Baykul (2010, pp.223–224) argued that rather than attempting to capture validity through a single definition, it is more appropriate to demonstrate the various sources of evidence supporting a test's validity. Therefore, it is essential to consider different types of validity rather than treating validity as a unitary concept. Validity is commonly classified into five main types: content, criterion-related, construct, face, and response validity (Alderson et al., 1995; Bachman, 1990; Brown, 1996; Heaton, 1988; Henning, 1987; Hughes, 2003)

Arguably, face validity is the first type of validity to be questioned when encountering an exam. As long as a test seems as if it measures what it is supposed to measure, one may say it has face validity (Hughes, 2003). Face validity refers to the extent to which a test appears appropriate and seems to measure the knowledge or abilities it claims to assess, reflecting its surface credibility and public acceptability (Ingram, 1977, p.18; Mousavi, 2002). One may undervalue face validity by labelling it as "less scientific"; however, it may have a potential impact on test results. Test takers may not take tests seriously if they appear irrelevant to the purpose, or they may perform to the best of their ability if the exam is face valid (Anderson, Clapham, and Wall, 1995). In this study, evaluators have looked for clues

regarding the surface level target-measure appropriateness and age-appropriateness of the tests.

Content validity refers to the degree to which the items on a test are representative and relevant samples of the content or abilities the test has been designed to measure (Brown and Hudson, 2002; Crocker and Algina, 1986). Aligning teaching and assessment criteria can promote skill development and enhance the washback effect of testing on instruction (Weir, 2005). Assessing content validity involves more than a basic evaluation of relation. A writing test may look like a writing test; however, its difficulty may be beyond the expected outcomes of that course or vice versa. It is important to consider students' language proficiency when preparing test items to make fair judgments (Cohen, 1994). Also, the outcomes it is designed to assess may be quite different from the course's outcomes. Hughes (2003) emphasizes that, within language testing, content validity is established when the test content accurately reflects a representative sample of the linguistic skills, structures, etc., it claims to measure. Hughes (2003) notes that content validity is crucial, as the higher a test's content validity, the more accurately it measures what it is intended to assess.

Content validity is determined by defining the scope of the target behavior, consulting a panel of experts, creating a framework to align the test with the target behavior, and summarizing the data obtained from this matching process (Crocker and Algina, 1986, p.218). Another approach to determining content validity is the use of a specification table based on expert judgment, in which experts are provided with the target behaviors and test items, asked to indicate whether each item measures the intended behavior with a yes/no response, and to rate each item on a scale from 1 ('strongly disagree') to 5 ('strongly agree'), with decisions made separately for each item based on their evaluations (Baykul, 2000, p.226). Alderson, Clapham, and Wall (1995) propose several methods for examining content validity, including comparing test content with the syllabus or specifications, gathering feedback from experts such as teachers, subject specialists, and applied linguists through questionnaires or interviews, and having expert judges evaluate test items and texts against specific criteria (p.193). In this study, evaluators are advised to use a specification table or item-outcome matching while deciding on the content validity. They have examined whether the related outcomes were represented in the test in accordance with their level of cognitive complexity.

Another important type of validity in exams is construct validity. Ebel and Frisbie (1991) define construct validation as the process of collecting evidence to ensure that a test

accurately measures an unobservable psychological trait and that the scores reflect their intended meaning. Similarly, Heaton (1988) explains that a language test demonstrates construct validity if it can measure particular characteristics in accordance with a theory of language behavior and learning. The term construct refers to any underlying ability hypothesized within a theory of language proficiency, and such theories may be confirmed, modified, or abandoned through testing (Bachman, 1990; Hughes, 2003). In this perspective, Henning (1987) points out that the aim of construct validation is to provide evidence that the theoretical constructs being measured are themselves valid.

Construct validity indicates how accurately a test assesses a construct by minimizing construct-irrelevant variance and emphasizing the theoretical components it is meant to capture (Haladyna and Downing, 2004). East (2016) warns that when test tasks incorporate irrelevant variables or fail to capture essential aspects of the construct, students may be prevented from demonstrating their actual abilities. Teachers can strengthen construct definition by selecting tasks aligned with the construct, providing clear task specifications, and applying criteria consistent with the test's aims (Luoma, 2004).

For example, if the intended outcome is "Students will be able to list phrases for booking in a recorded dialogue," it would be inappropriate to test this outcome by asking students to write definitions of the related phrases in the audio. The construct in this case would involve listening-related sub-skills such as identifying specific information, recognizing formulaic language in an audio, distinguishing main ideas from details, and identifying lexical sets or thematic vocabulary. Similarly, in a speaking task like "Students can express what they like or dislike," the scoring rubric should include fluency, accuracy, and vocabulary, but not comprehension, spelling, or punctuation. This illustrates why the careful selection of rubric items is crucial for ensuring construct validity.

According to Luoma (2004), there are three main approaches to defining the speaking construct in assessment. The linguistic approach focuses on language forms such as vocabulary, grammar, pronunciation, and fluency. The communicative approach emphasizes how effectively learners can use the skills and strategies required by test activities. Finally, the task-based approach highlights learners' ability to successfully engage with and complete test tasks.

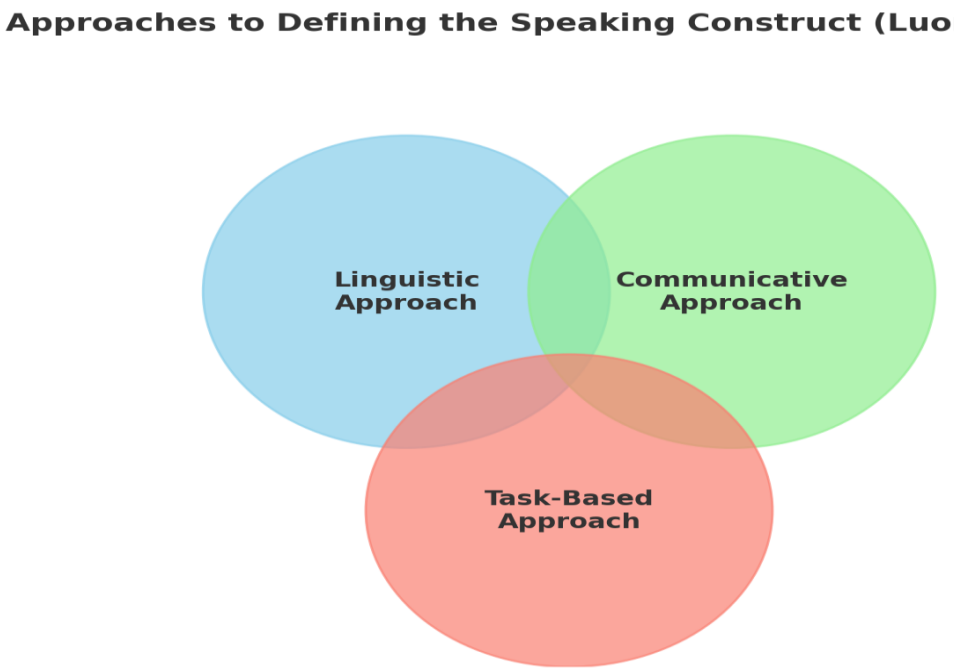


Figure 1. Approaches to Defining the Speaking Construct (Luoma, 2004)

Crocker and Algina (1986, pp. 371–374) mentions criterion-related validity. Criterion-related validity has two types: concurrent validity and predictive validity. Concurrent validity is about whether the test results can be supported by another existing criterion measure, such as a standardized test (Alderson, Clapham and Wall, 1995; Brown, 2004; Davies, Brown, Elder, Hill, Lumley and McNamara, 1999; Fulcher and Davidson, 2007; Murphy and Davidshofer, 2004). Concurrent validity practices can be highly useful when an appropriate criterion is available (Kılıç, 2015). With respect to the specific outcomes/grades targeted in this study, no standardized tests measuring the same outcomes were available; therefore, concurrent validity analysis was not applicable. Predictive validity is related to how well a test score can predict future success (Bachman, 1990; Brown, 2004; Hughes, 2003; Underhill, 1987). Predictive validity turns out to be significant in the case of placement exams, language aptitude tests or admission tests (Brown, 2004). The tests the study focuses on were not intended for these objectives, and there are not any achievement tests or standadized tests featuring language skills at the end of the year for comparison. Thus, predictive validity is also excluded in the study.

Another type of validity is response validity. Response validity concerns whether students respond in the intended way, yet factors such as unclear instructions or unfamiliar test formats may prevent them from demonstrating their actual ability (Henning 1987). Response

validity (response process evidence) requires collecting data on how students actually interpret and respond to test tasks (e.g., think-aloud protocols, interviews, eye-tracking, or analyzing cognitive strategies). In order to reduce unfamiliarity related problems the researcher preferred tenth grade second speaking tests as a sample and included some of the response validity related concepts in the checklist in broad terms under the umbrella term "construct validity". Because collecting detailed data on test-takers' cognitive processes across such a broad sample would have been impractical, this study did not investigate response validity. The construct validity argument presented here should be considered incomplete in this regard.

To sum up, for a high school speaking test to be face valid, one can expect speaking tasks that are familiar to the students, such as role-playing in real-life scenarios. As for the content validity, one may expect to see tasks requiring a similar level of competence to the outcomes taught in lessons. Also, one may not expect to observe abilities like memorization, but look for inclusion of speaking activities like pair dialogues or other types of communicative activities that are related to lesson outcomes as a sign of construct validity.

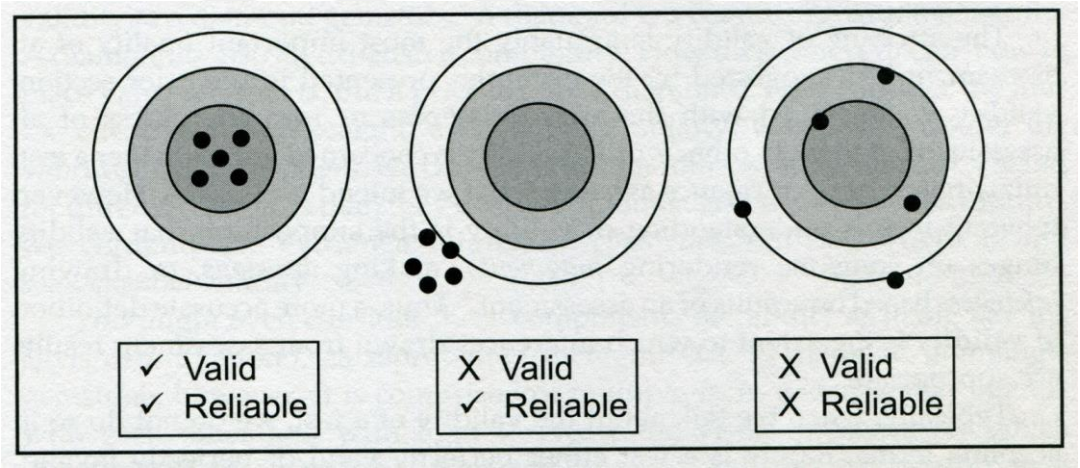


Figure 2. Validity and Reliability (Gareis and Grant, 2008)

2.2.2. Reliability in Speaking Tests

Reliability refers to the extent to which the results obtained from a measurement instrument are free from random errors (Turgut, 1990). A reliable exam is expected to give similar results when administered more than once in a short time. If a test cannot measure something consistently, it is very likely measuring it inaccurately, which means that an unreliable test

is also invalid (Anderson, Clapham, & Wall, 1995). However, a test may produce stable results but still fail to measure the intended construct, in which case it is reliable but not valid. Reliability can, to some extent, be strengthened through the use of clear tasks and instructions, multiple tasks or testing occasions (Green, 2014), well-defined criteria (Brown, 2004), and trained raters who follow standardization procedures (Luoma, 2004; Green, 2014; Brown, 2004). Hughes (2003) also points out that effective testing requires clarity and fairness, which in practice can be achieved by designing unambiguous tasks, providing familiar formats under uniform conditions, using detailed rubrics to keep objectivity, applying proper rating procedures such as multiple raters, and collecting enough samples of behavior to reduce error.

Reliability is generally divided into test reliability and rater reliability. Rater reliability can be further classified into intra-rater reliability, which refers to the consistency of an assessor's own judgments over time, and inter-rater reliability, which means the extent to which different raters award similar scores to the same performance (Jonsson and Svingby, 2007; Gampper, 2013). Intra-rater reliability is achieved when a rater gives consistent scores for the same student across occasions, including when re-evaluating a recorded performance, while inter-rater reliability is achieved when multiple raters produce comparable scores in the same context. Well-structured rubrics are often seen as important for supporting inter-rater agreement (Fulcher, 2003). Because rater issues are one of the most common sources of unreliability, this topic will be covered at in more detail later under Rater Related Issues section.

Brown and Abeywickrama (2010) list four main sources of unreliability: student-related issues, rater issues, test administration, and the test itself. Student-related factors include things such as fatigue, anxiety, health, motivation, or psychological state, all of which can lower the stability of performance. For example, Cohen (1994) mentions motivation, physical health, psychological mood, and test-taking experience as individual factors that may affect reliability. Administrative and situational conditions also play a role. Douglas (2014) argues that environmental factors such as comfortable seating, suitable temperature, enough lighting, and freedom from distracting noise are needed to create conditions where students can perform at their best. Hughes (2003) adds that scoring should be done in a quiet and well-lit environment and that tests should be carried out in places free of interruption. Similarly, Cohen (1994) identifies situational factors such as lighting, noise, comfort, and heating as variables that can influence reliability. Administrative considerations also matter,

as unclear instructions, inconsistent timing across students, and teachers' attitudes can all affect student performance.

Human judgment is another important source of error in testing. Rater bias, whether influenced by cultural background, prior knowledge, cognitive traits, first language, age, gender, or other factors, can have an effect on scoring outcomes (Bachman, 1990; Wang, 2010; Luoma, 2004; McNamara, 1996). Motivation, communication style, and prior experience as a rater or interlocutor can also play a role, while distraction and tiredness during the scoring process may interfere with consistent judgment. These issues are often hard to notice, which makes them particularly difficult to prevent. Engelhard (1994) and Lumley and McNamara (1995) point out that bias can appear as a systematic tendency of raters to give consistently higher or lower scores to particular groups of examinees or to certain tasks. Such tendencies add error to the measurement process and directly weaken the reliability of the test.

Test-related factors are also significant. Task difficulty has been found to influence students' performance in terms of accuracy, fluency, and complexity (Mehnert, 1998; O'Sullivan & Weir, 2002; Skehan, 1996; Skehan & Foster, 1997; Wigglesworth, 1997). To avoid problems with reliability, Douglas (2014) suggests that tests should be set at an appropriate level for the students: neither too difficult nor too easy. The length of the test is another concern. O'Sullivan (2013) stresses that young learners, in particular, should not be given long test tasks, as they lack the maturity to stay focused for extended periods. On the other hand Bachman (1990) argues that a five-minute speaking test is usually less reliable than a thirty-minute one, since longer assessments both reduce the impact of bias and provide a more representative sample of language performance. In this sense, test length directly affects the stability of scores and the adequacy of sampling. Cohen (1994) further explains that test factors such as the extent of sampling, clarity of instructions, and the use of familiar formats also contribute to reliability.

All of these studies show that reliability is many-sided and shaped by student characteristics, rater behavior, test administration, and the design of the test itself. Brown (1996) summed up these concerns in his list of potential sources of error variance (see Figure 3), which shows the wide range of factors that may affect the consistency of test results. While some of these influences can be controlled through careful test design, detailed rubrics, and standardized administration procedures, others, especially those tied to human judgment and individual test-taker differences, are much harder to remove completely. Therefore, reliability should

be understood as relative rather than absolute, with the practical aim being to reduce error and keep consistency as high as possible within the realities of educational assessment.

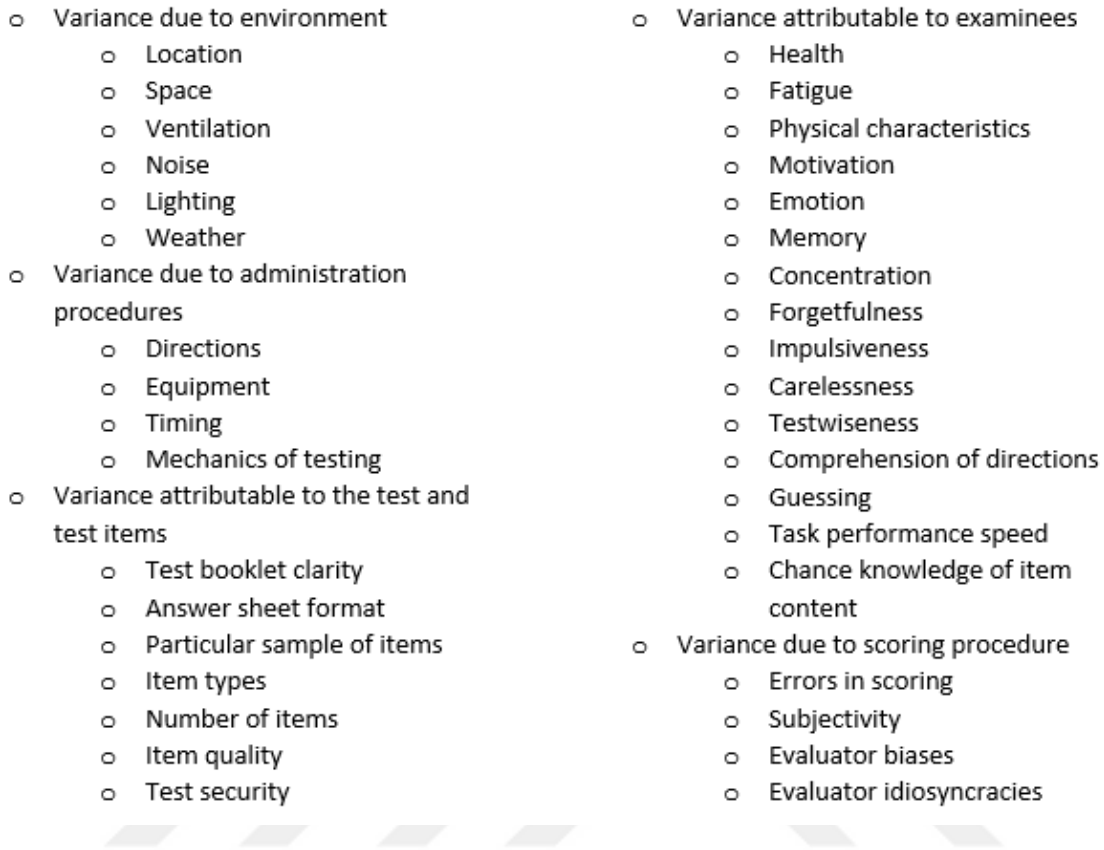


Figure 3. Checklist for Potential Sources of Error Variance (Brown, 1996, p.189)

Several methods that can be applied to assess and improve the reliability of a test (Weir, 2005) each of which helps ensure consistent measurement in testing. Some of them are:

- Test-retest method (Administering the same test to the same individuals under identical conditions at different times, and examining differences between scores.)
- Parallel (Equivalent) test method (Administering another test of similar characteristics alongside the target test, then examining the differences and similarities between results.)
- Split-half method (Dividing a test into two halves based on content and difficulty, and comparing the results obtained from both halves.)
- Multiple raters (scorers) (Having multiple raters score the same test paper, and examining the differences between their scores.)

- Intra-rater (Repeated scoring) (Having the same rater score the same test sample at different times, at least twice, and examining the differences between these scoring results.)
- External rater method (Having an educator other than the course instructor score the test.)
- Use of rubrics (scoring criteria) in evaluation.

In the current study, teachers were asked whether they used above mentioned methods, whether they had any precautions regarding the application of the test (environment, execution of the test, test takers' anxiety) and whether they used any type of rubric. The evaluators made use of the information provided by the teachers and the test itself, and evaluated five reliability related components, namely, clarity of instructions, clarity of the test content, rubric, rater reliability, and conduct.

2.3. Critical Issues in Oral Language Testing

2.3.1 Rubric Related Issues

Definitions of Rubrics and Rating Scales

The term rating scale is often used interchangeably with scoring rubric or proficiency scale (Fulcher, 2014; Fulcher and Davidson, 2007). Rubrics and scales are designed to capture the complex nature of language ability in structured ways. Underhill (1987) describes a rating scale as "a series of short descriptions of different levels of language ability," intended to guide assessors by showing what a typical learner at each level can do. Similarly, Harmer (2004) defines assessment scales as tools that specify scores for productive skills through "pre-defined descriptions of performance." In this sense, rating scales function as operational definitions of the construct being tested (Fulcher, 2003) and are inherently linked to the test's purpose, its construct definition, the tasks, and the criteria used in scoring (Luoma, 2004).

Rubrics share these functions but emphasize performance criteria more explicitly. They break down a task into predetermined categories and specify levels of achievement for each

(Stevens and Levi, 2005). Each level, from excellent to poor, is defined by a descriptor that typically outlines linguistic features such as vocabulary, syntax, or fluency, as well as the tasks and functions learners are expected to perform (Davies et al., 1999; Fulcher, 2010; Wigglesworth and Frost, 2017). As Fulcher (2014) explains, the purpose of a rubric or rating scale is to guide the rating process by aligning expectations with observable outcomes.

Functions of Rubrics in Speaking Assessment

Rubrics are considered effective tools for maximizing objectivity in the evaluation of productive skills (Brookhart, 2013). They provide a common framework that helps reduce subjectivity by making expectations clear. In addition to improving objectivity, rubrics offer practical benefits such as saving time, providing detailed feedback, and supporting the evaluation of teaching processes (Brookhart, 2013; Reddy and Andrade, 2010; Stevens and Levi, 2005). Their use can also strengthen validity, especially construct validity, since descriptors help check whether the intended criteria are actually met (Moskal, 2000).

Importantly, rubrics connect tasks with evaluation criteria. As Weir (2005) argues, tasks should not be considered separately from the criteria that will be applied to the performances they produce. This highlights the central role rubrics play in ensuring that what is being assessed corresponds meaningfully with what the task was designed to elicit.

Challenges in Application

Despite their benefits, rubrics present several challenges in practice. Because ensuring objectivity is critical for both reliability and validity (Pineda, 2014), developing and selecting an appropriate rubric is often demanding for teachers. Even when standardized rubrics are available, interpretation and application may vary significantly across raters, leading to notable scoring variation (Lumley, 2002). This variation shows the importance of providing detailed descriptors that can minimize ambiguity in scoring.

The difficulty of rubric use is particularly visible in speaking assessments. Rating speaking skills has consistently been identified as one of the most challenging tasks for teachers (Bachman and Palmer, 1996; Luoma, 2004). The literature frequently reports that different raters give different scores to the same performance on the same task (Ahmadi and Sadeghi, 2016; Brown, Iwashita and McNamara, 2005; Chalhoub-Deville and Wigglesworth, 2005; Han, 2016). McNamara (1996) further stresses that such inconsistencies can stem not only from the raters themselves but also from the design of tasks and rubrics.

The Turkish context also offers examples of the importance of rubrics. At a university in Türkiye, Höl (2010) reported that students experienced anxiety and low proficiency in speaking tests however teachers emphasized that assessment was feasible if clear rubrics were used. This finding highlights that rubrics not only shape rating practices but can also affect perceptions of fairness and manageability in speaking assessments.

Rater Cognition and Mental Rubrics

While rubrics are designed to promote consistency, rater cognition complicates their use. Rater cognition refers to the conscious or unconscious mental processes raters deal with while evaluating performances (Bejar, 2012; Davis, 2012; Myford, 2012; Han, 2016). These processes involve observing, interpreting, recalling, and weighing information. And of course, there is room for subjectivity in all of these.

Mental rubrics are a product of rater cognition. Campbell (1993) defines these as unwritten, experience-based frameworks raters construct in their minds, shaped by prior conceptions of oral proficiency and the criteria given in rubrics. Mental rubrics guide raters during scoring

but may not align perfectly with the written rubric, and they often exist despite training (Han, 2016). Research has shown that raters sometimes apply additional criteria not explicitly stated in the rubric, such as the quality or maturity of ideas, and interpret existing criteria differently (Brown, 2000). This can lead to substantial inconsistency in scoring, particularly in speaking tasks where performance is multidimensional.

Rubrics, Training, and Reliability

Training has been proposed as a way to minimize inconsistency and improve rubric use. Kim (2015), for example, demonstrated that novice and developing raters significantly improved their understanding and application of rubrics after repeated training sessions and personal feedback, whereas expert raters already showed desirable rating behaviors. However, training alone cannot entirely eliminate variability, especially when mental rubrics or personal judgments override rubric descriptors (Orr, 2002).

To address this problem, double or multiple raters and calibration sessions are recommended. Multiple raters help counterbalance individual biases, while calibration sessions allow raters to align their interpretations of rubric descriptors (Weir, 2005). Such measures are particularly important in speaking assessments, where the dual role of examiner and interlocutor complicates scoring. Double marking has been described as indispensable for enhancing reliability (Weir, 2005).

Bridging Rubrics and Raters

Ultimately, rubrics and raters are inseparable in speaking assessment. Rubrics lay out the structure and criteria, but their effectiveness depends on how raters interpret and apply them. Differences in rater judgment often emerge not from the absence of rubrics but from variability in their use and the influence of mental rubrics. Therefore, rubric-related issues cannot be fully understood without considering the human element of assessment. Ensuring consistency and fairness, therefore, requires not only well-designed rubrics with detailed descriptors but also systematic training, calibration, and the use of multiple raters. This

interdependence between rubrics and raters forms a key link in the literature and necessitates a closer examination of rater-related issues, which will be the focus of the following chapter.

2.3.2. Rater Related Issues

Rater Reliability

Raters are often identified as the main source of inconsistency in oral test scores because their judgments are inherently subjective, which makes rater reliability a central issue in speaking assessment (Bachman, 1990; Davis, 2012). As Luoma (2004) notes, scoring is shaped by the interaction among raters, criteria, and performances, and this interdependence underlines the necessity of rater training, well-designed scales, and standard-setting to enhance reliability. Yet, despite these measures, raters remain vulnerable to sources of variation such as fatigue, distraction, or misinterpretation of criteria. Underhill (1987) argued that the use of multiple assessors can increase reliability, and Bachman and Palmer (1996) suggested that two or three raters constitute an ideal arrangement.

Brown (2004) categorizes rater reliability into two primary types: inter-rater reliability and intra-rater reliability. Inter-rater reliability refers to the expectation that test-takers should obtain the same score regardless of which rater evaluates their performance (Fulcher, 2014). Where such consistency cannot be achieved, Luoma (2004) recommends either more comprehensive training for raters or clearer, more precise scoring criteria. Intra-rater reliability, on the other hand, concerns the stability of scores given by the same rater over time. In this regard, raters are expected to maintain internal consistency and to assign the same scores to identical performances across different occasions (Brown, 2004).

Leniency, Severity, and Measurement Error

One of the most widely discussed issues in speaking assessment is the effect of rater leniency and severity. Guilford (1954) described these tendencies as stable traits that raters naturally bring into any scoring situation, regardless of the task or examinee. Wolfe and McVay (2012) argued that rater severity or leniency constitutes a form of rater error, as test-takers receive

scores that are consistently higher or lower than their performances actually merit. Lumley and McNamara (1995) similarly found that even extensive rater training cannot completely eliminate these differences, as judgments continue to be influenced by prior experiences, biases, habits, and beliefs. Fulcher (2003) also emphasized that the outcome of an assessment may depend not only on the test-taker's linguistic competence but equally on the rater's severity or leniency.

Several researchers have highlighted the consequences of such tendencies. Wolfe (2004) explained that some raters are simply too lenient or too harsh in grading because of their personal characteristics. Myford and Wolfe (2004) went further, warning that leniency can constitute a serious measurement error, especially when student rankings are based on these scores. Congdon and McQueen (2000) similarly defined "rater stringency or severity" as the likelihood of a rater with elevated expectations assigning lower grades compared to other raters, a phenomenon that has persisted in testing for decades.

Rater Variability and Subjectivity

Ensuring rater objectivity is a major challenge in oral assessments. Çopur (2002) emphasized that scorers must demonstrate stability in evaluating the same test and maintain consistency across time and settings, yet as long as human judgment is involved, subjectivity is unavoidable. To mitigate this issue, assigning multiple raters to each performance has been widely recommended. Höl (2018), for example, noted that by comparing and correlating the scores given by multiple raters, it becomes possible to determine whether evaluations are consistent.

Rater subjectivity is an inherent factor in assessment and persists even with experienced evaluators. Shohamy (1983) and Fulcher (2003) observed that the judgments of highly trained raters often differ substantially from those of their peers. McNamara (1996) concluded that variability among raters is inherent, making it an unavoidable aspect of assessment. Acknowledging this variability and addressing it by employing multiple raters and relying on mean scores is therefore considered a reasonable approach.

2.3.3. Other Issues

External Influences on Scoring

Beyond rater traits such as leniency and severity, a number of subtle external factors can influence speaking scores. Wang Haizhen (2008) and Lane and Sone (2006) reported that variables such as the rater's tone of voice, body language, intonation, use of pauses, reactions when students fail to understand a question, attentiveness while listening, psychological state during scoring, expectations about test duration, and even the order of examinees can all affect judgments. For example, test-takers assessed at the beginning or end of a session may be disadvantaged or advantaged simply due to timing effects. Many of these influences are difficult to detect, making full objectivity in speaking assessment particularly challenging. In the present study, these types of variables were not analyzed, since investigating them would have required direct observation of all testing sessions across twelve schools, which was not feasible.

The relationship between rubrics and raters underlines how much human judgment shapes the outcome of speaking assessments. Even with clear criteria, raters' personal tendencies and interpretations can affect scores, which makes training and calibration essential for reliability. Yet, rater influence is only one side of the picture, as learners' own factors also play a key role and naturally lead to the discussion of test taker factors.

Test Taker Related Factors

Oral examinations often present unique challenges for learners, particularly because performance is judged in real time. Unlike written tests, there is little opportunity to revise or correct responses, which increases the psychological burden. Scholars have pointed out that it is essential to place candidates at ease from the very beginning so that they can reveal their true ability (Hughes, 2003). If the testing environment is poorly planned or tense, students may struggle to produce language naturally, even when their competence is adequate. Thus, the atmosphere and administration of the test become as important as the linguistic demands themselves.

Another frequent concern is the impact of test anxiety on oral performance. Speaking tests, while designed to provide learners with the best chance to demonstrate communicative skills (Bachman, 1990), often create stress and pressure that inhibit full performance. Research indicates that second language anxiety can significantly reduce fluency and confidence (Woodrow, 2006), and in some cases even prevent learners from engaging meaningfully in the task. Anxiety acts as a filter that interferes with output, masking actual ability and leading to results that do not reflect the learner's true proficiency (Heaton, 1988; Huerta-Macías, 1995). For this reason, oral testing remains one of the most affectively loaded forms of assessment, demanding careful attention to test design and administration.

Beyond anxiety, other learner-related factors also shape test performance. Familiarity with the test format influences how comfortable and effective students feel in demonstrating their abilities. For example, Şallı-Çopur (2002) found that many learners preferred individual interviews over other formats, suggesting that perceptions of fairness and comfort can directly affect outcomes. Motivation, willingness to communicate, and even personality traits such as introversion or extroversion further interact with performance in oral tests. A student who is well-prepared linguistically may still produce limited responses if shy or hesitant, while a more outgoing peer might take greater risks and appear more fluent.

In this study, these factors were not directly measured, but teachers were asked to report on the precautions they take to reduce stress and create a supportive test atmosphere. While not central to the analysis, these factors were indirectly analyzed through teacher input, which is a limitation in the study.

Task Design and Test Method Issues

The design of tasks and the selection of test methods are central to the fairness and effectiveness of oral assessments. Speaking tests often rely on certain task types such as interviews, role plays, narrations, and information-gap activities, all of which target key elements of oral proficiency, including fluency, accuracy, vocabulary, grammar, pronunciation, and interactional ability (Brown and Abeywickrama, 2010; Collie, 1998; Hughes, 2003). Such tasks are expected to provide opportunities for learners to use language for meaningful communication in specific contexts (Bachman and Palmer, 1996). However, when test content is limited to a narrow set of task formats or lacks authenticity, it risks

under-representing the construct of speaking and instead assessing test-taking strategies (Fulcher, 2003; Hughes, 2003; Luoma, 2004).

A key consideration in task design is the type of interaction the test requires. Individual interviews allow for flexible questioning and controlled scoring but may create power imbalances, authenticity concerns, and logistical challenges (Luoma, 2004; Weir, 2005). In contrast, paired and group tasks promote more natural communication, reduce anxiety, and support the development of interactional competence. Yet they also raise concerns about group composition and the difficulty of separating individual contributions (Brooks, 2009; May, 2009; Weir, 2005). Similarly, direct face-to-face tests are valued for authenticity, while indirect tests based on recordings are easier to standardize and less stressful but lack spontaneity and interactive features (Hughes, 2003; Luoma, 2004). Brown (2004) offers a practical classification, ranging from imitative and intensive tasks such as repetition or drills, to more extensive tasks such as monologues and presentations. Each type brings its own strengths in highlighting aspects of proficiency and its own challenges in scoring and administration.

Another challenge lies in balancing control, authenticity, and practicality. Informational and interactional tasks, such as descriptions, explanations, interviews, and discussions, vary in the degree to which they resemble real-world language use (Bygate, 1987; Luoma, 2004). Structured tasks may support reliability but restrict spontaneity, while open-ended tasks allow for richer language samples at the cost of scoring consistency. Planning time also emerges as a critical factor, particularly for lower-proficiency learners. Research suggests that when students are given time to plan the content and language of their responses, they can more effectively draw on their linguistic knowledge and produce more accurate and meaningful output (Skehan, 1996; Yuan and Ellis, 2003). Without such support, the demands of real-time formulation may cause hesitation or breakdowns in communication, affecting both fluency and accuracy.

All in all, task design in speaking assessment requires balancing authenticity, control, and practicality. It is important to keep in mind that tasks are not basic exercises, but goal-directed activities that shape what learners can demonstrate (Bachman and Palmer, 1996; Luoma, 2004). Because even minor variations in task type, interactional format, or preparation time can lead to significant differences in learner performance, careful selection and design are essential if tasks are to provide valid, fair, and representative measures of oral proficiency (Bachman, 1990; Fulcher, 2003).

In the context of the present research, task and method considerations are directly relevant because the speaking exams analyzed across different high school types in Sakarya were developed by classroom teachers rather than by large-scale testing bodies. This means that choices regarding task type, interaction format, and administration process are likely to influence the psychometric properties under investigation. For instance, an overreliance on interviews may raise concerns about authenticity, while paired tasks can complicate the reliability of scoring due to interactional variability. Memorization, for example, is not a speaking skill, yet some speaking tasks in the tests may solely depend on it. Similarly, the absence of planning time or the limited use of task variety may affect both the construct coverage and the degree of anxiety experienced by test takers. Looking at task design and test method issues helps put the validity and reliability findings of this study into context, since these factors reflect the practical realities of classroom-based speaking assessment.

2.4 Previous Studies Focusing on Speaking Tests

Speaking tests have been widely researched in the field of language assessment, applied linguistics and educational measurement due to their complex and challenging features. Previous studies focused on several issues, such as task design, technology-supported assessment and CEFR; however, the following review mostly concentrates on issues related to the psychometric properties of speaking tests.

Local Studies

Validity problems and skill coverage

Several studies show that speaking is often under-represented in exams. Ösken (1999) found that a final exam at a Turkish university focused mainly on vocabulary and grammar while ignoring writing, listening, and speaking, which created a content validity problem. Similarly, Cesur (2009) reported that standardized achievement tests had acceptable content and face validity but weak construct validity, again because speaking skills were not included enough. A larger study by Uzun and Kılıçkaya (2020) on national transition exams showed that test items matched the curriculum and coursebooks, but the four skills were not balanced

and speaking was neglected. At the school level, Sarıçoban (2011) examined mentor teachers' exams and found that they tested grammar and reading only, sometimes even including topics that had not been taught, leaving out productive skills altogether. These findings show that speaking has often been overlooked, which limits both content and construct coverage.

Reliability and rater issues

A number of studies focus on the consistency of scoring in speaking assessments. Ekmekçi (2016) compared native and non-native teachers' ratings and found high inter-rater reliability overall, with differences appearing mainly in pronunciation. Polat (2020) examined expert graders and showed that some scored too harshly and others too leniently, raising fairness concerns. İnan (2019) looked at how raters actually think while scoring and found that they often rely on their own internal criteria and biases in addition to the rubric. These studies suggest that while reliability can be reached, differences in rater behaviour remain an important issue in speaking assessment.

Test administration and practical issues

Other research highlights problems with the way speaking tests are carried out. Özdemir (2018) studied speaking exams in Anatolian high schools and found a lack of standardization, heavy dependence on teacher planning, and concerns about score inflation, even though teachers and students had generally positive attitudes toward the tests. Höl (2018) found that time limits in university speaking exams were seen as damaging both validity and reliability. These examples show that the way speaking exams are organized, including time, procedures, and atmosphere, affects how fairly and effectively they measure student ability.

Teachers' assessment literacy

Many studies underline the need for stronger teacher preparation in assessing speaking. İleri (2019) reported that both pre-service and in-service training give very little attention to productive skills, leaving teachers without enough knowledge of rubrics or reliability practices. Mede and Atay (2017) also showed that university prep school teachers felt confident about testing grammar and vocabulary but not about validity, reliability, or speaking assessment. Ballıdağ and İnan Karagül (2021) found that teachers across state schools judged their training in all areas of testing as weak and said they needed further basic training. In a similar line, Genç, Çalışkan, and Yüksel (2020) showed that teaching experience, graduation type, or prior training made little difference in teachers' knowledge about speaking assessment. More recently, Ak and Bozyiğit (2024) found that teachers on online communities often requested ready-made skill-based exams, again pointing to limited assessment literacy.

Test design and task formats

Some studies address the kinds of tasks used in exams. Akçay, Tunagür, and Karabulut (2020) analyzed exams in grades from five to eight and found an overuse of multiple-choice grammar and reading items, with little attention to speaking or listening. Sarıçoban (2011) also showed how exams often lacked any productive tasks. Where speaking is assessed, studies on raters (İnan, 2019; Polat, 2020) indicate that task type (individual or paired, open or structured) can affect how performance is judged. Özdemir (2018) also noted concerns such as score inflation in high school tests. These findings suggest that exam format, task choice, and scoring practice are closely connected.

Curriculum and policy alignment

Research also points to gaps between policy goals and classroom practice. Arslan and Üçok Atasoy (2020) found that even though MoNE policies stress skill-based and communicative assessment, teachers mostly used grammar- and vocabulary-focused written tests. Çimen (2022) showed that while teachers agreed with the purpose of the curriculum, their exams did not reflect its aims, relying on traditional formats instead of integrated or authentic tasks.

Yeşilçınar and Kartal (2020) found the same pattern with young learners: teachers said they valued speaking assessment but often avoided it due to lack of training and support. These studies highlight a consistent gap between national policy and classroom practice.

International Studies

Validity, reliability, and fairness in large-scale speaking tests

A number of studies examine how major speaking tests address the core qualities of validity, reliability, and fairness. Zhang and Elder (2009), in their review of the College English Test–Spoken English Test in China, reported that the test benefits from a structured, interactive format and standardized administration, which support reliability. Features such as double rating, examiner training, and consistent procedures also aim to ensure fairness. At the same time, the authors noted areas for improvement, including vague descriptors in the rating scales and limited interaction in group tasks, which raise concerns for both validity and reliability. Similar issues were observed by Koizumi (2022) in Japanese secondary schools, where classroom-based speaking assessments often showed limited task variety and low rater reliability, largely due to insufficient teacher training and the complex nature of speaking assessment. Koizumi calls for improved teacher preparation and flexible models of standardization to enhance assessment quality, to improve validity.

Rater-related studies

Several studies highlight the role of raters in shaping the reliability of speaking assessment. Kim (2015) compared new, developing, and experienced raters scoring ESL speaking performances across three sessions. Findings showed that experienced and well-trained raters produced more consistent scores, while less experienced raters varied more. The study confirms that rater training is essential for producing fair and reliable scores. Similarly, Galaczi (2008) examined paired speaking tests and found that cooperation between test takers affected scores: pairs who interacted smoothly tended to receive higher marks. While this supports the validity of interactional competence as a construct, it also raises fairness concerns, as unequal participation can mask individual ability. Galaczi suggested refining

rating scales, strengthening rater training, and including multiple task types to balance reliability and fairness.

Test-taker factors in speaking assessment

Research also shows how learner-related factors can influence speaking test performance and challenge score interpretations. Huang, Hung, and Hong (2016) investigated integrated speaking tasks and found that while language proficiency was the strongest predictor of test performance, topical knowledge and test anxiety also had significant effects. This indicates that scores may reflect not only language ability but also construct-irrelevant factors. Their findings underline the importance of considering test-taker characteristics such as prior knowledge and emotional state during test design and validation.

Task design and construct representation

Task type and design are other factors to affect what is measured in speaking assessments. Kyle, Crossley, and McNamara (2016) studied TOEFL iBT speaking tasks and found that independent tasks tend to reflect a single construct, while integrated tasks show variation across sub-types. This suggests that task format plays a crucial role in shaping construct validity and that integrated tasks may capture multiple skills at once. Galaczi (2008) also demonstrated how task interaction influences outcomes, as peer cooperation in paired tasks affects both scores and perceptions of fairness. These studies emphasize that careful task design is key to balancing validity and reliability.

Technology in speaking assessment

More recent studies have turned to technology-based speaking assessments. Getman (2020) analyzed the TOEFL Primary Speaking test with young learners from 11 countries. Results indicated that tasks and raters performed consistently across age groups and that learners themselves accounted for most of the score variance, which points to strong reliability. Xu,

Jones, Laxton, and Galaczi (2021) explored automated scoring in an online English speaking test. Their study showed that the automated system achieved high internal consistency, but it tended to be slightly more lenient than human raters, especially with lower-level speakers. These findings suggest that while automated scoring can support reliable measurement, differences from human judgment must be carefully monitored.

CHAPTER III

METHODOLOGY

3.1.Research Design

The research employed a qualitative case study design, as it aimed to investigate the psychometric qualities of classroom-based speaking tests in a specific context (Creswell, 2013; Merriam, 1998; Yin, 2014). Within this design, data were primarily collected through expert judgments, which provided systematic insights into the validity and reliability of the tests. The expert evaluation checklist, functioning as a structured observation form, was used to guide experts' evaluations in a consistent and theory-driven manner. This approach was chosen because case studies allow a close, detailed look at a specific situation. This study focuses on how speaking assessments are evaluated in Turkish high schools, bringing together theory and rich, context-based information.

3.2. Participants and Setting

The study was conducted in Sakarya Province of Türkiye, which is roughly located between Istanbul and Ankara, two major cities in Türkiye. It has a population of over one million citizens, and it stands out with its high youth population of 227.176 students between the ages of 5-17, according to the statistics of Sakarya Provincial Directorate of Education.

The population of the study consists of the English as a foreign language speaking tests in 103 high schools. Due to time and budget limitations and with the aim of achieving representativeness and comparability(Teddlie and Yu, 2007), a purposeful sampling strategy was employed. The maximum variation sampling, which aims to choose samples that show a wide range of characteristics related to a particular topic(Cohen, Manion and Morrison, 2018), was prefered in order to ensure representation of different types of schools. The sample consists of twelve different high schools(~%12 of total number of high schools in Sakarya province) including three anatolian high schools, three science high schools, three vocational high schools and three Imam Hatip (Religious Vocational) high schools.

The above mentioned four types of high schools are the most common types of high schools in the Turkish education system. Students graduating from the lower secondary school after the 8th grade go on to their education in high school. According to their national exam results, students may continue their education in "project high schools". While almost all of the science high schools are project schools, there are very few project schools in other school types.

Basic information regarding the school types and EFL classes (Board of Education, 2025):

Anatolian high schools are non-vocational high schools which prioritize teaching general academic subjects such as maths, chemistry, geography and history. From 9th grade to 12th grade, students have four hours of EFL classes every week.

Science high schools mostly focus on subjects like maths, geometry, physics, biology and chemistry. From 9th grade to 12th grade, students have four hours of EFL classes every week.

Vocational high schools offer students the fundamental skills required in the field. They offer various programs from fashion design to machinery, and core subjects such as maths, biology and foreign languages are also taught. Students have four hours of EFL classes in the 9th grade, and from 10th grade to 12th grade, students have only two hours of EFL classes every week.

Apart from the core academic subjects, Imam Hatip high schools deliver religious education such as Arabic, tafsir(interpretation of the Quran) and fiqh(principles of Islamic jurisprudence). Students have five hours of EFL classes in the 9th grade, and from 10th grade to 12th grade, students have only two hours of EFL classes every week.

Following the sampling process, the speaking tests were collected from the chosen schools. All selected tests are 1st-semester, 2nd skill-based English exams administered to 10th-grade students. This selection was made to reduce variability caused by external factors such as students' unfamiliarity with exam format, end-of-year fatigue, and teacher-student rapport. Each exam was analyzed by experts, including all related assessment tools(rubrics, instructions, scoring sheets, etc.). To maintain anonymity, all information related to schools, students, and teachers was removed before the review.

Four experts, all holding doctoral degrees, participated in the evaluation process. Each expert was a faculty member in English Language Teaching (ELT) departments at universities in Türkiye, with established academic experience in language assessment and teacher

education. They were selected on the basis of their advanced qualifications and expertise in ELT, ensuring that the evaluation process reflected informed professional judgment and met the methodological standards required for test validation studies. The number of experts was also considered sufficient for this type of validation study, as a minimum of three to five experts for establishing content validity and ensuring stable agreement metrics is recommended (Lynn, 1986; Polit and Beck, 2006).

3.3. Data Collection Tools and Procedures

Ethical Considerations

The data collection process was designed to ensure transparency, authenticity, and ethical appropriateness, in line with research standards(Creswell and Creswell, 2018). As required by the Sakarya University Institute of Educational Sciences, the researcher applied to the ethical committee for ethical approval of the study. Afterwards, permission from the provincial educational directorate is granted to collect data from the schools. The voluntary nature of the study and withdrawal rights were explained to teachers prior to data collection, and their signed consent forms were taken.

To protect participant confidentiality, all identifying information from the tests was anonymized. Each school was assigned a unique code (e.g., H01, H02, etc.), and all related materials were labelled accordingly before being distributed to the expert reviewers in line with the British Educational Research Association (BERA) (2018).

Data Collection Method

The researcher personally visited each participating school in the Sakarya province to establish rapport with teachers and collect the relevant documents firsthand, an effective technique that allows questions to be clarified and also unclear or incomplete answers can be followed up (Fraenkel, Wallen, and Hyun, 2012).

Each teacher was asked to complete a Speaking Exam Information Form, which provided contextual data about the structure and implementation of the assessment. In nearly all participating schools, sample exam materials and rubrics, used for scoring student performance, were collected. The materials included exam tasks, instructions, and any other documentation used during the assessment process.

The anonymized materials were shared with four field experts working in EFL departments in various universities across Türkiye. Delivering courses in English Language Teacher Education Programs and having at least a doctoral degree were the prerequisites while determining the experts. The experts were contacted via email and given access to the exam documents through a secure cloud storage system. They were offered multiple formats for reviewing and scoring the materials, including WORD/PDF files, Google Forms, and printed hard copies.

Rater Training and Evaluation Procedure

In order to ensure methodological rigor, expert raters were provided with detailed guidance before completing the evaluation checklists. This process functioned as a form of rater training, familiarizing the raters with the evaluation criteria, the scope of the tasks, and the expectations of the study. The guideline emphasized minimizing the effects of rater cognition, including tendencies toward leniency or severity, by clarifying evaluation procedures and descriptors in advance. Such preparatory steps are frequently recommended in language assessment research to strengthen inter-rater reliability and reduce subjectivity (Fulcher, 2003; Luoma, 2004).

The evaluation process was built around clearly defined criteria addressing face validity, content validity, construct validity, and reliability. For each dimension, raters were provided with operational definitions supported by established literature (e.g., Brown and Hudson, 2002; Ebel and Frisbie, 1991; Hughes, 2003; Turgut, 1990). This alignment with recognized frameworks aimed to ensure that raters were not only consistent but also theoretically grounded in their judgments. For example, the guideline directed raters to consider item–outcome alignment when judging content validity and to assess the appropriateness of rubrics and task specifications when considering construct validity.

To further enhance scoring consistency, the descriptors “Yes,” “Partially,” and “No” were explicitly defined and accompanied by sample statements. This descriptor system functioned as a rubric-like scale, providing raters with anchors to interpret criteria in a structured way. By including examples of acceptable and unacceptable cases, the descriptors were intended to reduce ambiguity and supported more reliable decision-making. The use of such structured categories mirrors best practices in rater training, where detailed descriptors are considered key to minimizing misinterpretation and promoting transparency in the evaluation process (Brookhart, 2013; Weir, 2005).

The curriculum documents and annual plans, provided to the experts, ensured that raters had access to relevant benchmarks when judging the representativeness and appropriateness of test tasks. Finally, the data included contextual safeguards to prevent bias, such as anonymizing schools (e.g., H01, H02). Taken together, these measures illustrate a systematic approach to rater training and evaluation, designed to maximize methodological soundness and to provide trustworthy judgments on the psychometric properties of the speaking tests under investigation.

Data Collection Instruments and Validation

The expert evaluation checklist (see Appendix 5), functioning as a structured observation form, was developed to guide experts in examining four main psychometric qualities of speaking tests: face validity, content validity, construct validity, and reliability. Each dimension is grounded in established frameworks in language testing. For example, face validity follows Hughes’ (2003) emphasis on surface appropriateness, while content validity draws on Brown and Hudson’s (2002) and Alderson, Clapham, and Wall’s (1995) notion of representativeness of outcomes and item–outcome alignment. Construct validity reflects Ebel and Frisbie’s (1991) definition of measuring intended psychological traits, further supported by Luoma’s (2009) call for clear task–construct alignment. Reliability items are linked to classical definitions of measurement stability (Brown, 2004; Turgut, 1990) and incorporate rater reliability concerns discussed by McNamara (1996) and Jonsson and Svingby (2007).

The descriptors (Yes, Partially, No, I don’t know) were provided to structure the observation process, ensuring that experts recorded their judgments with clarity and consistency

(Fulcher, 2010). The accompanying guidelines also referenced best practices such as specification tables and bias reduction strategies, aligning with recommendations by Crocker and Algina (1986) and East (2016). In this way, the form was both theoretically anchored and practically applicable, serving as a structured observation instrument consistent with methodological standards in test validation research.

Alongside the expert checklist, a Test Information Form (see Appendix 4) was collected from teachers to provide detailed information about each exam. Since many factors affecting validity and reliability are not always visible in the test paper itself, this form ensured that experts had access to contextual details such as scope, methods, and administration practices when making their evaluations.

The form was designed by the researcher to align with the major dimensions of test validation identified in the literature. The development process adhered to principles of effective instrument design as emphasized in measurement literature (Fraenkel, Wallen, and Hyun, 2012). Questions on exam scope and targeted outcomes supported judgments about content validity by showing how well test tasks reflected curricular goals (Alderson et al., 1995; Brown and Hudson, 2002). Items on methods, duration, and preparation procedures related to construct validity and practicality, since task design and time allocation influence how well a test measures intended skills (Ebel and Frisbie, 1991; Luoma, 2004). Information on administration, such as clarity of instructions, precautions during testing, and classroom conditions, was linked to reliability, consistent with studies stressing situational influences on performance (Cohen, 1994; Hughes, 2003). Finally, teacher-reported use of rubrics or multiple raters directly linked to the literature on ensuring scoring reliability (Jonsson and Svingby, 2007; Weir, 2005).

By providing this contextual information, the form complemented the expert checklist and helped reduce the risk of misinterpretation. In line with validation frameworks (Bachman and Palmer, 1996; Weir, 2005), it ensured that expert judgments were based not only on test content but also on how the exams were designed and implemented in practice.

In order to better evaluate the sample tests gathered from the schools, experts benefited from the Test Information Form. The form includes both closed-ended and open-ended items aimed at collecting data on:

- Exam coverage (units covered),
- Assessment methods used (dialogue, presentation, etc.) in detail,

- Pre-exam preparations (announcements, task design, etc.),
- Preparation and performance time allocated to students,
- Environmental or procedural measures taken during the exam,
- Scoring process and methods used (e.g., rubric-based evaluation),
- Specific reliability strategies employed (e.g., inter-rater reliability, repeated scoring, parallel forms).

The development of both the expert evaluation checklist and the test information form involved multiple rounds of expert consultation to strengthen their validity. Two specialists were consulted: one from the field of measurement and evaluation and another from English as a Foreign Language (EFL) field. With one expert, two consultation rounds were conducted, and with the other, three rounds were held. These repeated rounds of discussion led to substantial refinements in both instruments, ensuring that the criteria and items aligned more closely with established frameworks in language testing (Bachman and Palmer, 1996; Hughes, 2003; Weir, 2005).

For the checklist, revisions included removing overlapping or ambiguous wording, clarifying items that lacked clear evaluation criteria, separating double-focused items into distinct entries, and merging items that assessed the same construct (e.g., different methods of estimating reliability). Similarly, the test information form was enhenced by introducing clearer question formats, replacing less common terms with more widely accepted ones, and structuring items to guide teachers step by step toward providing more detailed information. These adjustments reflect best practices in rater guidance and instrument design, where clarity, focus, and usability are essential for reducing bias and enhancing reliability (Fulcher and Davidson, 2007; Jonsson and Svingby, 2007).

Most of the revisions stemmed directly from expert feedback, with only minor modifications made after the pilot administration. Through these systematic validation steps, the instruments were refined to achieve both theoretical soundness and practical applicability. The next section describes the pilot study, which served as an additional step to confirm the effective use of the final versions.

Pilotting

A pilot administration was conducted in two schools with high numbers of students, where data were collected from three teachers. This piloting phase was essential in identifying possible logistical or procedural issues (Dörnyei, 2007). In the pilot study, data collection was not conducted face-to-face; instead, teachers completed the forms independently and submitted them afterwards. During this process, teachers were asked to complete the Test Information Form and to provide sample exams together with any supplementary materials. While the forms were returned after the exam sessions, the materials provided a first insight into how the instruments would function in practice.

The analysis was carried out by the researcher, focusing both on the teachers' responses and on how these responses could be interpreted alongside the expert evaluation checklist. This dual approach allowed the researcher to identify not only incomplete or ambiguous answers but also potential challenges that experts might face when evaluating the data in the main study. The pilot revealed that some teachers tended to provide brief or superficial responses, occasionally leaving sections incomplete or lacking detail, which created uncertainty for interpretation.

In response to these findings, several adjustments were made before the main study. The form was revised to include expanded response spaces and more explicit prompts, helping teachers provide more precise and detailed information. In addition, it was decided that data collection would be carried out face-to-face in the main study. This adjustment was intended to give teachers the opportunity to ask questions about technical terms, encourage them to reflect more carefully on their answers, and reassure them that their responses would not affect their professional standing. Overall, the pilot study served its purpose of identifying potential weaknesses in the instruments and procedures, leading to improvements that aimed to ensure clarity, completeness, and reliability in the subsequent data collection.

3.4. Data Analysis

The data gathered from the expert evaluations were analyzed quantitatively through descriptive and inferential statistical methods with the aim of determining the extent to which

the speaking tests met key psychometric standards, namely face validity, content validity, construct validity and reliability. Data normality is calculated via Shapiro-Wilk test. Descriptive statistics (frequencies, percentages, means) summarized the expert evaluation data. Inferential statistics, including Fleiss' Kappa, Intraclass Correlation Coefficients, and the Kruskal-Wallis test, were used to assess inter-rater reliability and look for significant differences among the school types.

The internal consistency was examined using Cronbach's alpha. Alternative coefficients, such as ordinal alpha or McDonald's omega, were also considered. However, these methods assume more complex data structures and are more beneficial when scales have a wider range of response categories. Since the present instrument uses a three-point response format (Yes–Partially–No) and aims to provide a commonly interpretable reliability estimate, Cronbach's alpha was selected as the most appropriate measure.

Inter-rater Reliability

Fleiss' Kappa was calculated across the 13 checklist items. Fleiss' Kappa is suitable for measuring agreement among more than two raters and was applied to determine the overall inter-rater reliability of the four expert evaluations. Interpretation of Kappa values followed the benchmarks established by Landis and Koch (1977), where values between .61–.80 indicate substantial agreement, values between .41–.60 represent moderate agreement, values between .21–.40 represent fair agreement, and values between .0–.20 represent slight agreement.

To comprehensively assess the consistency of the expert panel, inter-rater reliability was analyzed using Intraclass Correlation Coefficients (ICC). A two-way random-effects model for absolute agreement (ICC [2,k]) was employed, as it effectively accounts for both rater and target variance. A key methodological advantage of ICC is its capacity to generate two distinct metrics: one for the reliability of a single rater's score (ICC [2,1]) and another for the reliability of the panel's average score (ICC [2,k]). This provides a nuanced understanding of measurement precision, indicating whether decisions can be based on an individual expert's judgment or if the collective consensus of the group is required for reliable measurement (Koo & Li, 2016; McGraw & Wong, 1996). Comparison of ICC interpretation benchmarks are displayed in Table 1.

Table 1

Comparison of ICC Interpretation Benchmarks

ICC Value Range	**Koo & Li (2016)**	***ICC Value Range***	**Cicchetti (1994)**
< 0.5	Poor	< 0.4	Poor
0.5 to < 0.75	Moderate	0.4 to < 0.6	Fair
0.75 to 0.9	Good	0.6 to < 0.75	Good
> 0.9	Excellent	≥ 0.75	Excellent

Comparative Analysis

Comparative Analysis. The Kruskal–Wallis H test was used to examine whether there were statistically significant differences in exam quality scores across the four school types. This non-parametric method was chosen because the data did not meet the assumptions of normality and homogeneity of variances (Field, 2013). The test allows for comparing more than two independent groups based on ranked data, making it appropriate for the ordinal scoring scale used in the expert evaluations (Pallant, 2016).

3.5. Methodological Findings

3.5.1. Internal Consistency: Cronbach's Alpha

The internal consistency was examined using Cronbach's alpha. Although the items were ordinal (Yes, Partially, No), Cronbach's alpha remains the most common and practical

reliability coefficient for multi-item Likert-type scales. It provides a direct indication of how consistently the items measure the same construct. For ordinal data, Cronbach's alpha is known to be a conservative estimate, meaning that the true reliability of the scale is often slightly higher than the reported value. This makes it a reliable and safe choice for assessing internal consistency.

Internal Consistency of Scale Items

As presented in Table 2, the item-total correlation analysis revealed that the majority of items were moderately to strongly correlated with the overall scale across all four raters ($r > .50$), confirming satisfactory internal consistency of the speaking assessment checklist. High correlations were particularly observed for Items 1, 2, 4, 5, 6, 10, and 13, indicating that these items contributed positively to the homogeneity of the scale. However, Items 7, 8, and 9 consistently showed weak correlations ($r < .30$) across raters, suggesting that they may not be fully aligned with the construct being measured. The "Cronbach's Alpha if Item Deleted" values also demonstrated that removing these low-correlating items would result in a slight increase in the overall reliability coefficients. These findings indicate that, while the checklist as a whole exhibits strong internal coherence, a few items might benefit from revision or clearer descriptors to ensure consistent interpretation by raters.

Table 2

Corrected Item–Total Correlations and Cronbach's Alpha if Item Deleted for All Raters

Item	Rater 1 CITC	Alpha if Deleted	Rater 2 CITC	Alpha if Deleted	Rater 3 CITC	Alpha if Deleted	Rater 4 CITC	Alpha if Deleted
Item 1	0.81	0.82	0.54	0.78	0.77	0.84	0.81	0.82
Item 2	0.78	0.83	0.32	0.79	0.70	0.85	0.75	0.83
Item 3	0.62	0.85	0.41	0.80	0.60	0.86	0.50	0.85

Item 4	0.69	0.84	0.28	0.82	0.73	0.83	0.64	0.84
Item 5	0.71	0.83	0.35	0.81	0.65	0.85	0.65	0.84
Item 6	0.59	0.86	0.44	0.79	0.61	0.86	0.63	0.84
Item 7	0.11	0.87	0.08	0.87	0.01	0.87	0.04	0.87
Item 8	0.26	0.86	0.09	0.88	0.15	0.87	0.12	0.87
Item 9	0.37	0.85	0.11	0.86	0.25	0.86	0.29	0.86
Item 10	0.79	0.83	0.61	0.78	0.83	0.82	0.82	0.82
Item 11	0.51	0.85	0.38	0.80	0.49	0.85	0.48	0.85
Item 12	0.42	0.86	-	-	0.09	0.87	-	-
Item 13	0.68	0.84	0.47	0.79	0.69	0.84	0.65	0.84

Note. Dashes indicate items automatically removed by SPSS due to zero variance.

According to conventional benchmarks, Cronbach's Alpha values below 0.70 indicate low reliability, those above 0.70 are considered acceptable, above 0.80 are good, and above 0.90 are excellent (Nunnally and Bernstein, 1994; Tavakol and Dennick, 2011).

Internal Consistency of Experts

The internal consistency of the raters' scores was examined using Cronbach's Alpha. The results indicated high internal consistency for Rater 1 (α = .894), Rater 3 (α = .872), and Rater 4 (α = .857), and moderate consistency for Rater 2 (α = .693). These findings suggest that the majority of raters applied the rubric consistently, while minor variability in Rater 2's scoring may indicate differences in interpretation or application of the criteria.

Table 3

Cronbach's Alpha Coefficients for Each Rater

Expert	Cronbach's Alpha	Alpha (Standardized)
Expert 1	.894	.889
Expert 2	.693	.653
Expert 3	.872	.855
Expert 4	.857	.849

3.5.2. Inter-Rater Reliability

The consistency of ratings among the four experts was assessed using multiple statistical measures to provide a solid evaluation of reliability. Fleiss' Kappa, Weighted Kappa and Intraclass Correlation Coefficients (ICC) were employed to determine the inter-rater reliability. Fleiss' Kappa, Weighted Kappa, and ICC together offer a comprehensive picture of inter-rater reliability. Fleiss' and Weighted Kappa capture how consistently experts categorize each item (e.g., Yes, Partially, No), while ICC (2,k) specifically evaluates the reliability of the average ratings across multiple raters. In this way, the kappa coefficients reflect agreement at the categorical level, whereas ICC (2,k) assesses the stability of the composite expert judgment, making these analyses complementary within the study's context.

As previously discussed, rater cognition and individual characteristics play a key role in the evaluation process. Therefore, examining the frequency distribution of expert responses may provide valuable insight into their rating tendencies.

Table 4

Frequency Distribution of Responses by Experts

Expert	**Yes**	**Partially**	**No**
Expert 1	40 (29.0%)	51 (37.0%)	47 (34.1%)

Expert 2	65 (47.1%)	44 (31.9%)	29 (21.0%)
Expert 3	43 (31.2%)	64 (46.4%)	31 (22.5%)
Expert 4	19 (13.8%)	66 (47.8%)	53 (38.4%)

An examination of the frequency distributions revealed noticeable differences in rater tendencies. Expert 2 displayed a clear tendency toward leniency, assigning the highest proportion of positive ("Yes") judgments (47%). In contrast, Expert 4 emerged as the most severe rater, offering the fewest "Yes" responses (14%) and the highest rate of negative ("No") ratings (38%). Such differences illustrate common rater-related patterns, including leniency, severity, which are frequently observed in performance assessments (Eckes, 2015; McNamara, 1996; Myford and Wolfe, 2004). To account for these variations and to evaluate the extent of consistency across raters, inter-rater reliability was examined using Fleiss' Kappa and Intraclass Correlation Coefficients (ICC).

3.5.3. Agreement Analysis with Fleiss' Kappa

Table 5 presents the results of the Fleiss' Kappa analysis conducted to examine the level of agreement among experts across different dimensions of the checklist. The table includes the number of valid ratings, observed and expected agreement values, the Fleiss' Kappa coefficients, and their interpretations based on Landis and Koch's (1977) classification. This provides an overview of how consistently the experts rated each dimension of the assessment tool.

Table 5

The results of the Fleiss' Kappa Analysis by Dimension.

Dimension	**Total Valid Ratings (N)**	**Po (Observed Agreement)**	**Pe (Expected Agreement)**	**Fleiss' Kappa (κ)**	**Interpretation***

Face Validity (1–2)	87	0.655	0.426	0.400	Fair agreement
Content Validity (3–4)	88	0.693	0.453	0.439	Moderate agreement
Construct Validity (5–8)	175	0.663	0.351	0.481	Moderate agreement
Reliability (9–13)	216	0.718	0.344	0.569	Moderate–substantial agreement
Overall (1–13)	**566**	**0.687**	**0.343**	**0.524**	**Moderate agreement**

Note: Interpretation based on Landis and Koch, 1977 $\kappa < .20$ = slight; .21–.40 = fair; .41–.60 = moderate; .61–.80 = substantial; > .80 = almost perfect.

The Fleiss' Kappa analysis revealed moderate to good levels of inter-rater agreement across the dimensions. Overall agreement among the four experts was $\kappa = .52$, indicating a moderate-to-good consistency (Landis & Koch, 1977). Among the sub-dimensions, Reliability showed the highest level of agreement ($\kappa = .57$), while Face and Content validity demonstrated moderate agreement ($\kappa \approx .40$–.44). Construct validity also showed moderate agreement ($\kappa = .48$). These findings suggest that while there is notable alignment among raters, some variability remains, especially in the evaluation of face and content aspects.

3.5.4. Agreement Analysis with Weighted Kappa

To further examine agreement between individual raters, pairwise Weighted Cohen's Kappa values were calculated using quadratic weights. In practice, for ordinal rating scales with three or more ordered options, quadratic weights are often preferred because they better represent the "distance" between categories (Cohen, 1968; Warrens, 2015). The results of the Weighted Kappa analysis are presented in Table 6.

Table 6

Weighted Kappa (Pairwise & Quadratic) – With 95% CIs

Rater Pair	N	Quadratic κ	95% CI	Label
Expert 1 – Expert 2	138	0.529	[0.414, 0.632]	Moderate
Expert 1 – Expert 3	139	0.520	[0.402, 0.623]	Moderate
Expert 1 – Expert 4	139	0.347	[0.203, 0.479]	Fair
Expert 2 – Expert 3	141	0.572	[0.461, 0.664]	Moderate
Expert 2 – Expert 4	142	0.355	[0.240, 0.471]	Fair
Expert 3 – Expert 4	142	0.511	[0.392, 0.615]	Moderate

The results showed that agreement ranged from fair to moderate across all rater pairs. Quadratic κ values ranged from .35 to .57, as expected for ordinal data. The strongest agreement was observed between Expert 2 and Expert 3 (quadratic $\kappa = .57$), while the lowest occurred between Expert 1 and Expert 4 (quadratic $\kappa = .35$). These results align with the overall Fleiss' Kappa (.52) and ICC findings, suggesting moderate reliability when ratings are considered collectively.

3.5.5. Score Reliability Analysis with Intraclass Correlation (ICC)

To complement the agreement analysis, Intraclass correlation coefficients were computed using a two-way random effects model, assessing both single and average rater consistency. This model was selected as both the raters and the exams were considered random samples from larger populations. The results, presented in Table 7, show that the single-rater ICC [ICC(2,1)] was 0.48 (95% CI = 0.38–0.58, $p < .001$), indicating moderate reliability. In contrast, the average-rater ICC [ICC(2,k)] was 0.79 (95% CI = 0.71–0.84, $p < .001$), reflecting good reliability when the ratings of all four experts were aggregated (Koo & Li, 2016). These results highlight that while individual rater agreement was only moderate, combining ratings across experts substantially improved reliability. This confirms that the

aggregate judgments of the expert panel provide a consistent and reliable measure for evaluation.

Table 7

Intraclass Correlation Coefficients for Inter-Rater Reliability

ICC Model	Description	ICC Value	95% CI	Interpretation
ICC(2,1)	Two-way random, single rater	.48	[.352, .472]	Moderate
ICC(2,k)	Two-way random, average raters	**.727**	[.662, .784]	**Good**

Note: Interpretation based on Koo & Li (2016): <.50 = Poor, .50-.75 = Moderate, .75-.90 = Good, >.90 = Excellent.

Table 8 displays the Intraclass Correlation Coefficients (ICC) calculated for each dimension of the checklist. The table includes both single-measure [ICC(2,1)] and average-measure [ICC(2,k)] values, along with their standard deviations, to show the consistency of expert ratings across dimensions.

Table 8

Intraclass Correlation Coefficients ICC by Dimension

Dimension	ICC(2,1) Mean (SD)	ICC(2,k) Mean (SD)
Face Validity	0.46 (0.06)	0.73 (0.04)
Content Validity	0.33 (0.10)	0.65 (0.10)
Construct Validity	0.29 (0.05)	0.61 (0.05)
Reliability	0.32 (0.19)	0.58 (0.18)

The ICC analysis revealed notable variation across dimensions. Face validity achieved the highest consistency among raters (ICC(2,1) = 0.46; ICC(2,k) = 0.73), reflecting moderate reliability at the single-rater level and good reliability when aggregated. Content validity showed lower but still acceptable agreement (ICC(2,1) = 0.33; ICC(2,k) = 0.65). Construct validity produced weaker values (ICC(2,1) = 0.29; ICC(2,k) = 0.61), indicating limited consistency. Reliability items were the most heterogeneous, with ICC(2,1) averaging 0.32 but with high variability (SD = 0.19), suggesting inconsistency across different aspects of reliability, though aggregated ratings improved to moderate levels (ICC(2,k) = 0.58).

3.6. Interpretation of the Methodological Findings

The reliability analyses revealed that both the checklist and the expert evaluations showed generally acceptable and stable levels of consistency. Cronbach's Alpha results indicated that the speaking assessment checklist had a coherent internal structure, meaning that most items worked well together to measure the same construct. Three raters (R1, R3, and R4) demonstrated high internal consistency ($\alpha > .85$), while Rater 2 showed a moderate level (α: .69). This suggests that, overall, the raters applied the checklist criteria in a consistent way, although one rater may have interpreted some descriptors slightly differently. Such small variations are quite expected in performance-based assessments, as raters naturally differ in how they perceive and prioritize specific aspects of student performance (Eckes, 2015; McNamara, 1996).

The item-level findings provided additional insight into the scale's functioning. Items 1, 2, 4, 5, 6, 10, and 13 appeared to be the most reliable, showing strong correlations with the overall scale. These items therefore supported the internal coherence of the instrument. On the other hand, Items 7, 8, and 9 showed weak correlations ($r < .30$), which may indicate that their wording or focus was not fully clear to all raters. The "alpha if item deleted" results also showed that removing these weaker items could slightly increase the overall reliability. In other words, the checklist worked effectively as a whole, but a few items might need clearer descriptions or examples to prevent different interpretations. Even so, Cronbach's Alpha values were all above the acceptable threshold of .70 (Nunnally and Bernstein, 1994; Tavakol and Dennick, 2011), which confirms that the scale had satisfactory internal consistency.

When the focus shifts from each rater's consistency to the level of agreement among raters, the results paint a realistic picture of how subjective judgments can vary in classroom-based assessment. The Fleiss' Kappa analysis showed a moderate overall agreement (κ: .52) among the four experts. The agreement was strongest for the Reliability dimension (κ: .57), while Face and Content validity produced slightly lower scores (around κ: .40–.44). This pattern is quite typical in language testing research because raters often find it easier to agree on observable aspects such as scoring clarity or reliability than on more abstract criteria like authenticity or representativeness (Fulcher, 2003; Myford and Wolfe, 2004).

Pairwise Weighted Kappa results supported this interpretation. The agreements between individual pairs of raters ranged from fair to moderate (κ: .35–.57). The highest consistency was found between Expert 2 and Expert 3, whereas the lowest was between Expert 1 and Expert 4. This finding was also consistent with the frequency distribution results, where Expert 2 tended to be more lenient and Expert 4 more severe. These differences once again highlight the influence of individual rating styles. Although detailed criteria and descriptors were provided to guide the experts' evaluations, slight differences in interpretation still emerged. This situation is also frequently reported in the literature, showing that even when raters receive common training or use the same rubric, variability in judgment can persist due to personal perceptions and cognitive differences (Eckes, 2015; Lumley and McNamara, 1995).

The Intraclass Correlation Coefficients (ICC) provided another perspective on the reliability of expert scores. The single-rater ICC(2,1) value (.48) indicated moderate reliability, but when the scores of all four raters were combined, the average ICC(2,k) rose to .73, showing good reliability (Koo and Li, 2016). This means that while a single rater's scores may vary moderately, combining the ratings of several raters gives a much more stable and trustworthy result. This finding supports the common recommendation in performance testing to use multiple raters to minimize individual bias and increase reliability (Bachman, 1990; Brown and Abeywickrama, 2019).

Across the dimensions, Face Validity showed the highest ICC values (ICC(2,k): .73), while Construct Validity and Reliability dimensions produced moderate but still acceptable levels (around ICC(2,k): .58–.61). This indicates that consistency was higher in more concrete aspects of the test, such as clarity and transparency, compared to more abstract aspects like construct representation.

To sum up, the reliability analyses showed that the expert evaluations were overall consistent, and the speaking test checklist functioned as a dependable measurement tool. Although some variation among raters was observed, particularly in interpreting certain items, the overall consistency remained within acceptable limits. Taken together, these results confirm that the combination of well-prepared checklists and multiple expert evaluations can provide reliable evidence about the quality of classroom-based speaking exams. In future applications, revising a few items and organizing short rater calibration sessions could further strengthen the reliability of the results and reduce potential interpretation differences among evaluators.

CHAPTER IV

FINDINGS

This chapter presents the results of the expert evaluations regarding the validity and reliability of the exams from 11 schools. The findings are presented in the following order: (1) General descriptive statistics, (2) Findings for each validity and reliability dimension, (3) School-based analysis, and (4) Varience among School Types.

4.1. General Descriptive Findings

An overview of the evaluations across all 11 schools and 13 items is presented below (Table 9). A total of 572 data points were analyzed (11 Schools* × 13 Items × 4 Experts).

*One school was excluded from the research, since no summative speaking tests were applied in the school, instead the teachers relied on inclass performance observation to assign a score for the students' speaking proficiency.

Table 9

Overall Descriptive Statistics for All Evaluations (no opinion excluded) (N=566)

Category	**Frequency (f)**	**Percentage (%)**
Yes	170	30.0
Partially	234	41.3
No	162	28.6
Total	**566**	**100**

Overall, experts' ratings averaged 2.01 (SD = 0.77) on the three-point scale, suggesting that judgments generally leaned toward "partially". Six "no opinion" responses were excluded.

4.2. Dimensional Analysis

The performance across the four main dimensions is summarized in Table 10.

Table 10

Comparative Performance of the Four Main Dimensions

Dimension	Number of Items	Mean Score (SD)	Rank
1. Face Validity	2	2.43 (0.68)	1
2. Content Validity	2	2.24 (0.61)	2
3. Reliability	5	1.91 (0.77)	3
4. Construct Validity	4	1.82 (0.78)	4

Note: SD = Standard Deviation

As shown, **Face Validity** received the highest average rating, while **Construct Validity** dimension received the lowest.

4.2.1. Face Validity Findings

Table 11 presents the descriptive results for the first two items (G1–G2) in the expert evaluation checklist. These items address the clarity of the test's alignment with the intended learning outcomes and the appropriateness of the exam tasks for the target age group. The percentages of "Yes," "Partially," and "No" responses, along with the standard deviation values, are displayed to illustrate the distribution of expert judgments.

Table 11

Face Validity Item Analysis

Item	Description	Yes (%)	Partially (%)	No (%)	SD %

G1	appears to be clearly related to the intended learning outcome.	47.73	40.91	11.4	19.3
G2	seems appropriate for the participants' age group (14–16).	58.14	32.56	9.3	24.4
Average		52,90	36,75	10,35	21.85

Item G2 (Age-appropriateness) was the highest-rated single item in this dimension.

4.2.2. Content Validity Findings

Table 12 presents the descriptive results for the two items (K1–K2) under the content validity dimension of the checklist. These items focus on the inclusion of learning outcomes within the exam content and the appropriateness of the assessment tool for the cognitive levels targeted by the outcomes. The table displays the percentages of "Yes," "Partially," and "No" responses, together with the standard deviation values, to summarize the distribution of expert evaluations.

Table 12

Content Validity Item Analysis

Item	**Description**	**Yes (%)**	**Partially (%)**	**No (%)**	**SD %**
K1	The learning outcomes specified within the scope of the exam are included in the assessment.	31.8	56.8	11.4	22.8
K2	The assessment tool is appropriate for the cognitive levels of the learning outcomes	34.1	59.1	6.8	26.1
Average		32.95%	57.95%	9.1%	24.45

Both content validity items showed a very similar profile, with a high proportion of "Partially" responses.

4.2.3. Construct Validity Findings

Table 13 presents the descriptive results for the four items (Y1–Y4) under the construct validity dimension of the checklist. These items examine whether the assessment tool accurately measures the intended learning outcomes, avoids assessing unintended skills or knowledge, includes measures to minimize potential bias, and employs a clear scoring rubric aligned with its purpose. The percentages of "Yes," "Partially," and "No" responses, together with the standard deviation values, are provided to summarize the distribution of expert evaluations.

Table 13

Construct Validity Item Analysis

Item	Description	Yes (%)	Partially (%)	No (%)	SD %
Y1	The assessment tool is designed to measure the intended learning outcomes.	41.9	51.2	7.0	23.3
Y2	The assessment tool does not require knowledge or skills beyond the intended learning outcomes.	50.0	25.0	25.0	14.4
Y3	The assessment practice includes measures to prevent possible bias.	0.0	31.8	68.2	34.1
Y4	A clear and detailed scoring rubric suitable for the purpose has been used in the assessment practice.	0.0	38.6	61.4	31.0
Average		22,98	36,65	40,40	25.7

This dimension shows the most variation. Items Y1 and Y2 performed reasonably well, while items Y3 (Bias prevention) and Y4 (Scoring rubric) were the two lowest-rated items in the entire study, pulling down the dimension's average.

4.2.4. Reliability Findings

Table 14 presents the descriptive results for the five items (Gv1–Gv5) under the reliability dimension of the checklist. These items address the clarity of test instructions and content, the appropriateness of scoring criteria, the assurance of rater reliability, and the suitability of the exam administration for its intended purpose. The percentages of “Yes,” “Partially,” and “No” responses, along with the standard deviation values, are displayed to summarize the distribution of expert evaluations.

Table 14

Reliability Item Analysis

Item	Description	Yes (%)	Partially (%)	No (%)	SD %
Gv1	The test has clear, precise, and understandable instructions.	25.0	36.4	38.6	7.3
Gv2	The content of the test is clear, explicit, and comprehensible.	61.4	34.1	4.6	28.4
Gv3	The criteria used for scoring are appropriate for the purpose and content.	12.2	65.9	22.0	28.6
Gv4	Rater reliability has been ensured.	0.0	6.8	93.2	51.9

Gv5	The stated exam administration is appropriate for the purpose and content.	27.9	60.5	11.6	24.9
Average		25,30	40,74	34,00	28.22

Item Gv2 (Content clarity) was a strength, while Gv4 (Inter-rater reliability) was a notable weakness.

4.3. School-Based Performance Analysis

Table 15 shows the ranking of schools based on their total mean scores from the expert evaluations. Each school is identified with a code number, and the table lists its mean score, standard deviation, and rank order. The ranking offers a clear overview of how the schools performed relative to one another according to the overall evaluation results.

Table 15

School Performance Ranking (Based on Total Mean Score)

School Code	Mean Score (SD)	Rank
10	2.33 (0.79)	1
12	2.25 (0.79)	2
3	2.23 (0.78)	3
6	2.21 (0.80)	4
2	2.13 (0.66)	5
5	2.08 (0.71)	6
11	2.07 (0.77)	7

1	1.76 (0.69)	8
8	1.75 (0.65)	9
4	1.67 (0.73)	10
7	1.62 (0.67)	11

There is a clear performance gap between the top schools (e.g., School 10, 12, 3) and the bottom schools (e.g., School 8, 4, 7). The scores range from a high of 2.33 (School 10) to a low of 1.62 (School 7).

4.4. Varience Among School Types

Prior to group comparisons, the normality of the averaged checklist scores was examined, and the data were found to deviate from normal distribution. Therefore, a non-parametric Kruskal–Wallis H test was employed to explore potential differences in expert evaluation scores among Science, Religious, Anatolian, and Vocational high schools.

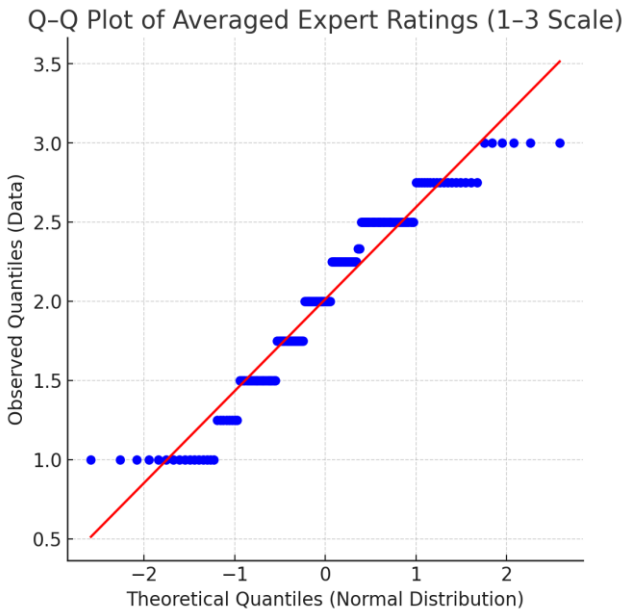


Figure 4. Q-Q plot of averaged expert ratings

Figure 4 showing the relationship between observed and expected normal values. The data points diverge from the diagonal reference line, indicating that the distribution of scores does not fully follow a normal pattern, in line with the Shapiro–Wilk test results.

The Shapiro–Wilk test indicated a significant deviation from normality, $W = 0.94$, $p < .001$, confirming that the averaged expert ratings were not normally distributed. Given the ordinal nature of the original checklist data (Yes = 3, Partially = 2, No = 1) and the non-normal distribution of the mean scores, the data were analyzed using non-parametric methods. Accordingly, the Kruskal–Wallis H test was applied instead of ANOVA to ensure an appropriate comparison across school types.

A Kruskal–Wallis H test was conducted to determine whether expert evaluation scores differed significantly among the four school types: Science (S), Religious (R), Anatolian (A), and Vocational (V) high schools. The analysis revealed a statistically significant difference among the groups, $H(3) = 27.99$, $p = .002$. As illustrated in Figure 4.4, the median values and score distributions varied across school types, indicating that the speaking assessment evaluations were not uniform across contexts and that certain school types demonstrated higher levels of alignment with the assessment criteria than others.

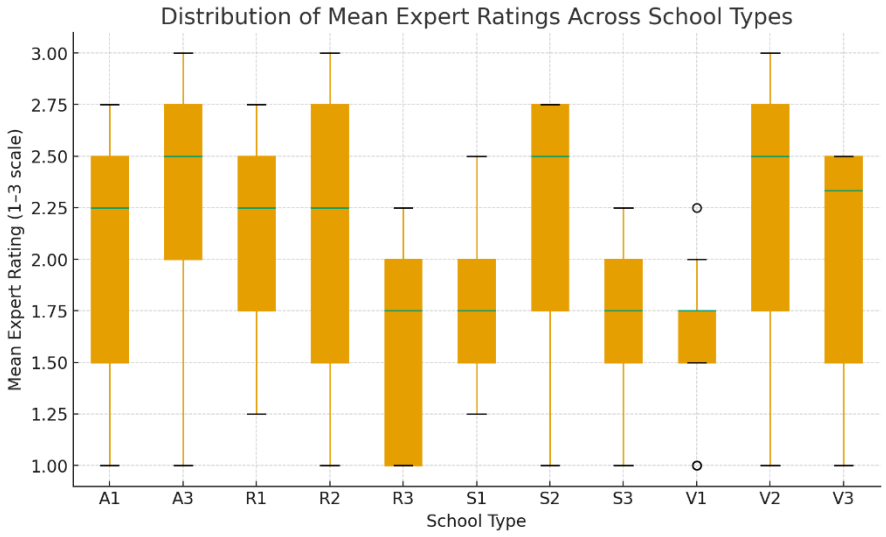


Figure 5. Boxplot of mean expert evaluation scores across four school types (Science, Religious, Anatolian, and Vocational)

The figure illustrates variations in medians and score distributions, consistent with the Kruskal–Wallis test results indicating significant differences among the groups.

The comparison of expert evaluation scores across the four school types revealed notable differences in how speaking assessments were conducted and rated. As shown in Figure 4.4,

Science high schools generally received higher mean ratings, suggesting stronger alignment with the assessment criteria and more consistent implementation of evaluation procedures. Anatolian and Religious high schools exhibited moderate median scores, indicating a balanced but slightly variable approach to speaking assessment. Vocational high schools, on the other hand, tended to have lower mean ratings and wider score spread, implying greater inconsistency in assessment practices. These differences suggest that contextual and institutional factors may influence how effectively speaking assessment standards are applied across school types.

The higher mean expert ratings observed in Science high schools indicate that the speaking tests administered in these institutions were perceived as more consistent with validity and reliability criteria compared to other school types. This suggests that test design and administration practices in Science high schools may demonstrate stronger adherence to established assessment principles.

CHAPTER V

DISCUSSION AND CONCLUSION

5.1. Overview of the Chapter

This chapter interprets the results of the expert evaluations on the speaking exams collected from twelve high schools in Sakarya. It focuses on understanding what the findings reveal about the validity and reliability of these exams and explores possible reasons behind the patterns observed. The discussion centers on four key dimensions: face, content, and construct validity, together with reliability, while also considering how they relate to one another in practice.

Both statistical results and expert insights are examined together to explain what these figures mean in the real testing context. Attention is given to the ways in which exam design, task selection, and rubric use may have influenced the psychometric qualities of the tests. Throughout the discussion, interpretations are supported by relevant theories and earlier research on language assessment to provide a balanced and evidence-based perspective.

5.2. Validity and Reliability

5.2.1 Face Validity

Face Validity received the highest mean score among all four dimensions (M: 2.43, SD: 0.68), indicating that the experts perceived the speaking exams as generally appropriate and visually credible in relation to their intended purposes. At the item level, G1 (“clearly related to the intended learning outcome”) reached 47.7 percent “Yes” and 40.9 percent “Partially,” while G2 (“appropriate for the participants’ age group”) obtained the strongest agreement, with 58.1 percent “Yes.” Overall, more than half of all expert judgments fell on the positive end of the scale, suggesting that the tasks were considered well matched to student profiles and curricular expectations. However, high face validity does not guarantee whole alignment, and perceived relevance can mask hidden design issues (Getman, 2020).

The relatively high ratings on G1 show that the tasks were perceived as visibly aligned with the stated objectives of the speaking component. Experts generally agreed that test prompts resembled classroom practices and reflected what students were expected to achieve in the curriculum. G2 received the most positive evaluations, indicating that task content and difficulty levels were age-appropriate and linguistically accessible for learners aged 14–16. The small proportion of "No" responses (below 10 percent in both items) supports the idea that the tasks were within learners' communicative reach.

This interpretation is reinforced by the statement from one teacher (H01): "I choose the topics among the ones students can talk about." This comment reflects teachers' awareness of learners' cognitive and linguistic limits and explains why age-appropriateness emerged as the most positively rated feature. It also suggests that face validity is not merely a matter of task design but a reflection of the teacher's intentional adaptation to student realities, which is also visible in a participant teacher's statement (H11) "We do not push the students so much. We choose topics they can perform"

These findings are consistent with earlier research in Turkish context. Ekmekçi (2016) and Özdemir (2018) both found that school-based English speaking exams often achieve high face validity due to teachers' familiarity with curricular goals and their tendency to design practical, familiar tasks. Similarly, Hughes (2003) emphasized that when test takers recognize exam tasks as meaningful and relevant, they are more likely to perceive the test as fair and meaningful. In the present context, the use of everyday and predictable topics appears to have strengthened this perception of validity.

The strong performance in face validity may also be linked to the clarity of the Ministry of National Education (MoNE) guidelines, which necessitates conducting speaking tests. Even though the procedures can be demanding or time-consuming, teachers appear to comply carefully, possibly because adherence is institutionally enforced.

Teachers' adherence to official assessment guidelines can be influenced by institutional and policy-driven factors. According to Ball, Maguire, and Braun (2012), policy enactment at the school level often involves teachers interpreting and operationalizing formal requirements in locally meaningful ways. In the context of language assessment, Alderson and Wall (1993) highlight how mandatory exams and official testing procedures can generate strong washback effects, shaping teachers' instructional and assessment practices.

This compliance ensures that certain surface features, such as task format, time limits, and question types, are standardized, producing tests that appear coherent and legitimate to external evaluators. In this sense, the high face-validity scores likely reflect teachers' adherence to official instructions. Consequently, while the exams are externally convincing and age-appropriate, they may still vary in depth and construct coverage, an issue explored further in the next sections.

5.2.2 Content Validity

Content validity was the second-highest dimension with a mean of 2.24 (SD: 0.61), yet the high number of "Partially" responses (57.9 percent on average) indicates that alignment between exam content and the targeted learning outcomes was only partial. Both items, K1 (coverage of intended outcomes) and K2 (appropriateness for cognitive levels), showed almost identical distributions: around one-third of "Yes" responses, more than half "Partially," and less than ten percent "No." These results suggest that experts recognized some consistency with the curriculum but also perceived missing or simplified elements within the examined speaking tests.

The item-level findings suggest that teachers tend to cover the visible or easily measurable parts of the curriculum rather than the full range of speaking outcomes. For K1, which measures how well the test includes the stated learning outcomes, experts noted that some teachers limited their assessment to a few units or familiar topics, leaving broader communicative skills untouched. For K2, which concerns the match between the assessment and the cognitive level of the intended outcomes, raters observed that several tasks were simplified to fit students' limited proficiency levels. This practice shows an effort to maintain fairness but also results in an underrepresentation of higher-level speaking outcomes.

Several test descriptions and expert comments help to clarify this pattern. Teacher H05 reported, *"I allow students to choose between two topics,"* reflecting a flexible but narrowed approach to content selection. Similarly, H07 wrote, *"Students draw a card and answer the question with one or two sentences."* However, the questions on those cards included items such as *"Translate: Biz Sakarya'da yaşarız – Jack reads newspaper"* and *"Talk about your relatives (names, ages, jobs),"* which reveal a deviation from the course outcomes. Expert evaluations confirmed these inconsistencies. Rater 1 (R1) noted for H05 that *"it was*

mentioned that the test covered Units 3 and 4, but while it's related to Unit 3, it has nothing to do with Unit 4." R1 also observed for H07 that *"it ended up being a testing instrument that only required memorization and recall."* Such remarks show that teachers often claim curricular alignment but end up designing tasks that assess memorized content instead of authentic communicative performance.

These results resonate strongly with earlier research on local speaking assessment practices. Özdemir (2018) and Ekmekçi (2016) both found that classroom speaking exams in Turkish high schools frequently reflect partial coverage of curricular outcomes due to time limitations and the absence of clear task models. According to Brown and Abeywickrama (2019), content validity is achieved when assessment tasks reflect the range of skills, functions, and topics outlined in the instructional goals, rather than emphasizing isolated linguistic elements. In the present study, the experts' partial evaluations confirm that teachers achieve topical overlap with the curriculum but not full representativeness of its cognitive and communicative demands.

The moderate content validity may be explained in part by contextual constraints and teacher adaptation. For instance, research in Turkish high schools indicates that English teachers often report limited weekly instructional time and challenges in covering all skills thoroughly (Dursun, Bedir and Gülcü, 2017). As a result, teachers may focus on fewer, more manageable outcomes or simplify tasks to ensure completion. Furthermore, when they aim for broader coverage, some teachers perceive the ministry-specified outcomes as overly demanding for students' proficiency levels, and thus feel compelled to adjust them downward, which was also visible in Ak and Bozyiğit's (2024) research. This leads to practical and age-appropriate but linguistically narrow exams. Consequently, while teachers succeed in maintaining visible content alignment with the curriculum, the assessments tend to prioritize feasibility and compliance over comprehensive content representation.

5.2.3 Construct Validity

Construct validity received the lowest mean score among all four dimensions (M = 1.82, SD = 0.78), indicating that the experts found the exams least satisfactory in representing the underlying construct of speaking ability. The four items (Y1–Y4) under this category varied considerably. While Y1 ("measures intended learning outcomes") and Y2 ("does not require

knowledge or skills beyond the outcomes") received moderate support with around half of the experts selecting "Yes" or "Partially," Y3 ("includes measures to prevent possible bias") and Y4 ("uses a clear, detailed scoring rubric") were among the weakest items in the entire checklist. Only 31.8 percent of raters selected "Partially" and none selected "Yes" for Y3, while Y4 received zero "Yes" and 38.6 percent "Partially," confirming the lack of standardized rubrics and bias-control procedures in the examined exams.

The item-level findings show that most speaking exams partially addressed the intended learning outcomes but through tasks that simplified complex communicative functions. Teachers often replaced open-ended, spontaneous speaking outcomes with limited or predictable prompts, leading to underrepresentation of the construct. For Y1, experts commented that several exams reduced speaking to sentence recall or memorized responses rather than real-time interaction. One evaluator noted for H03 that *"the student will read from memory. There will be no speaking practice,"* illustrating how task design limited the measurement of actual speaking ability.

Y2 received slightly better ratings, reflecting teachers' efforts to keep tasks within students' proficiency range. The sharp declines in Y3 and Y4 reveal more systemic weaknesses. None of the examined schools employed procedures such as double marking or external rating, and all rubrics lacked descriptors. Experts repeatedly noted the absence of detailed scoring rubrics, commenting for instance on H10 that *"no detailed rubric, no descriptors, no external raters."* This absence of structured rating criteria undermines the interpretive meaning of scores, as judgments become dependent on teachers' intuitive impressions rather than shared standards.

None of the teachers across the eleven schools provided a detailed rubric with discriptors. The rubrics attached to exam materials were highly general, typically dividing scores equally among broad categories such as *fluency 20p, accuracy 20p, pronunciation 20p, vocabulary 20p,* and *content 20p* (H06) or *pronunciation & accuracy 20p, fluency 20p, body language & eye contact 20p, extended vocabulary 20p, timing 20p* (H08). While these indicate a certain level of awareness of common assessment dimensions, they lack behavioral descriptors and therefore do not guide raters on what constitutes each performance level. One teacher (H07) explicitly admitted, *"I make an overall evaluation. Although I do not allocate points separately according to some dimesions, I care about things like pronunciation."* Such statements confirm that teachers tend to evaluate holistically based on general impressions, an approach that limits construct representation and scoring reliability.

The weak construct validity observed in this study echoes the concerns raised in previous national and international research. In the Turkish context, Polat (2020) and İnan (2019) highlighted persistent problems related to rater subjectivity, insufficiently defined rubrics, and scoring criteria that merely list assessment components without providing behavioral descriptors. Likewise, Özdemir (2018) and Ekmekçi (2016) reported that high school speaking exams often depend on teachers' intuitive judgments rather than standardized rating tools, leading to limited comparability and interpretive ambiguity.

Similar patterns have been identified internationally. Fulcher (2010) and McNamara (1996) noted that construct validity in speaking assessment is often weakened when tasks elicit memorized responses or when scoring lacks analytic grounding. Kim (2015) and Koizumi (2022) also identified vague task constructs and low inter-rater reliability as persistent obstacles to valid measurement. At a theoretical level, Bachman and Palmer (1996) emphasized that a test's validity depends on whether it elicits the full range of language ability components which are organizational, pragmatic, and strategic, rather than isolated linguistic forms. The near-complete absence of bias-prevention mechanisms observed in this study further reinforces these long-standing concerns.

The observed weak construct validity may largely stem from contextual and practical constraints. For instance, English teachers in Turkish high schools often report restricted weekly instruction time, leaving limited opportunities for preparation, administration, and scoring of comprehensive speaking assessments (Dursun, et al., 2017). Under such conditions, teachers may favour simpler, more manageable tasks to ensure completion even when these reduce the breadth of the construct being measured. Measures designed to reduce bias, such as double-marking, audio-recording of tests, or systematic cross-checking are often omitted, probably by the same reason that they impose additional workload and require institutional coordination that schools may not routinely support. Furthermore, the development and use of analytic rubrics with clearly defined descriptors demand a level of assessment literacy that not all teachers feel they possess (Ballıdağ and İnan Karagül, 2021; Genç, Çalışkan and Yüksel, 2020; İleri, 2019); hence many resort to holistic scoring based on an overall impression of students' language ability. While such an approach is efficient from a practical viewpoint, it compromises thorough construct representation and transparency. Therefore, although the speaking exams may fulfil procedural requirements and appear superficially authentic, they fall short of capturing the complex, interactive and strategic dimension of speaking ability as envisaged in communicative assessment theory.

5.2.4 Reliability

The reliability dimension received an overall mean score of 1.91 (SD: 0.77), placing it third among the four main dimensions. The five items (Gv1–Gv5) within this category showed uneven results. While Gv2 ("clarity of content") achieved relatively strong support with 61.4 percent "Yes" responses, Gv4 ("rater reliability ensured") was the lowest-rated item in the entire checklist which received zero "Yes" responses, but received 93.2 percent "No." Other items, such as Gv1 ("clarity of instructions") and Gv5 ("appropriateness of exam administration"), hovered around the "Partially" range, reflecting modest but inconsistent reliability across schools.

Experts consistently rated the clarity of test instructions and content (Gv1 and Gv2) positively. This can be explained by teachers' everyday classroom experience, giving clear directions is a fundamental part of teaching, and even when assessments differ from regular lessons, instructional clarity tends to remain strong. However, issues emerged in other reliability-related items.

For Gv3 ("appropriateness of scoring criteria"), experts noted several inconsistencies between what the tasks required and what the scoring criteria included. For example, one evaluator said for H03 and H06 that *"the use of present and past tense is required, but grammar is not included in the scoring,"* showing a mismatch between test content and assessment focus. Conversely, in H04, grammar was overemphasized, as the expert wrote, *"the teacher stated they focused on grammar. A focus specific to speaking skills was not observed."*

The weakest item, Gv4, concerned rater reliability, and all experts agreed that there were no measures such as double scoring, external raters, or calibration. This absence was common across all schools and is likely due to workload and scheduling constraints. Finally, Gv5 ("appropriateness of exam administration") revealed procedural inconsistencies. Some teachers provided the questions in advance (H05, H03), allowing memorization and recitation. Experts noted that this practice undermined reliability by creating unequal preparation conditions, as seen in comments such as *"Students will speak from memory. They will prepare in advance for the question they want. This is not appropriate."*

Test information forms confirm that exams were often designed for practicality rather than standardization. One teacher (H11) reported, *"Say five sentences using 'used to'."* While such prompts are clear and directly linked to the grammar related outcomes in the curriculum, they do not reflect the nature of speaking skills or authenticity. Another teacher stated, *"We told students to say five old habits, if they couldn't do it, we asked questions like 'What is your father's name?'"* (H11). These examples illustrate well-intentioned adjustments to help students succeed but also show how spontaneous decisions during administration may compromise fairness and reliability. Furthermore, none of the tests involved external raters, recordings, or formal moderation processes, indicating that reliability was largely dependent on individual teacher judgment.

These findings are in line with observations in both national and international literature. McNamara (1996) and Fulcher (2010) emphasized that reliability in speaking assessment is often the first dimension to be compromised in classroom contexts, largely due to limited time, workload, and the absence of systematic moderation. In the Turkish context, Özdemir (2018) similarly noted that teachers seldom employ multiple raters or structured scoring procedures for practical reasons. Brown and Abeywickrama (2019) also reminded that procedural consistency—not just task clarity—plays a decisive role in ensuring reliability in performance-based tests.

Further evidence supports these patterns. Zhang and Elder (2009) cautioned against the use of loosely defined rubrics and vague scoring criteria, both of which were apparent in the present study. Previous research has also shown that some level of subjectivity is inevitable in rater judgments, and even experienced assessors may vary substantially in their scoring (Fulcher, 2003; Shohamy, 1983). Considering that each exam in this study was evaluated by a single teacher, it is reasonable to assume that individual differences in judgment further affected reliability. As Polat (2020) pointed out, the rating process itself can be demanding and context-sensitive, even under ideal conditions. While certain procedural aspects, such as instruction clarity, were evaluated positively, the lack of consistent rater training and standardized procedures limited reliability overall. This mixed picture aligns with the findings of Ballıdağ and İnan Karagül (2021) and Mede and Atay (2017), who reported insufficient professional preparation in maintaining inter-rater consistency in school-based speaking assessments.

The reliability results reflect a common tension between practicality and standardization in classroom assessment. Although teachers are accustomed to giving clear explanations and

ensuring comprehension, conducting structured speaking assessments with consistent scoring and controlled conditions seem much harder. Time limitations, heavy teaching loads, and the absence of institutional frameworks for rater training all may have contributed to weak rater reliability (Brown, 2005; Fulcher, 2010; Kremmel and Harding, 2020).

Some teachers also seemed to prioritize student comfort and perceived fairness by simplifying testing procedures, sharing questions beforehand, or allowing flexible response formats. Although such adaptations may foster a supportive classroom atmosphere and reduce student anxiety, they simultaneously limit the comparability and replicability of assessment outcomes (Fulcher, 2010; Shohamy, 2001).

In short, reliability in these speaking exams appears procedural rather than psychometric: teachers ensured clarity and fairness in delivery, but the lack of standardized scoring and moderation prevented the results from achieving consistency across raters or contexts. This finding suggests that improving reliability in school-based speaking assessment would require institutional support, practical training, and structured tools that go beyond individual teacher effort.

5.3. Variance Among School Types

A Kruskal–Wallis H test was conducted to examine whether expert evaluation scores differed significantly among the four school types: Science, Anatolian, Religious, and Vocational high schools. The analysis revealed a statistically significant difference, $H(3)$: 27.99, p: .002, indicating that the quality of speaking assessments varied notably across contexts regarding validity and reliability. Descriptive data showed that Science high schools received the highest mean scores (M: 2.33), followed by Anatolian (M: ≈ 2.21) and Religious high schools (M: ≈ 2.08), while Vocational high schools recorded the lowest (M: 1.62). This imbalance suggests that test quality and adherence to validity and reliability principles were not evenly distributed among school types.

The significant variation across school types reflects broader educational and structural differences in Türkiye's high-school system. Science high schools admit academically high-achieving students and operate as project schools. These factors likely contribute to their higher expert ratings. Anatolian and Religious high schools displayed moderate results, consistent with their student profile. Vocational high schools, by contrast, showed weaker

alignment with validity and reliability criteria. Their limited English-instruction hours, only two per week, and general focus being on the vocational subjects may restrict opportunities for oral practice and test preparation, which in turn may affect the exam quality.

In schools with fewer English hours, exams tended to rely on short, predictable prompts and memorized responses, suggesting that teachers calibrated task difficulty to their students' limited exposure to English.

The present findings are consistent with earlier national studies reporting contextual disparities in classroom assessment quality. Özdemir (2018) found that schools with stronger academic profiles and more instructional hours tend to conduct speaking assessments that are both broader in content and more aligned with communicative-language-teaching principles. Internationally, Fulcher (2010) and Brown and Abeywickrama (2019) observed that institutional context and teacher experience play decisive roles in determining assessment quality, especially when resources and training opportunities differ. The significant variance found here therefore reflects not merely random fluctuation but a systematic pattern rooted in contextual realities.

The observed differences among school types underline how institutional resources, student profiles, and curriculum time jointly shape the validity and reliability of classroom-based speaking exams. Teachers appear to make pragmatic adjustments, such as simplifying tasks, shortening responses, or reducing assessment scope to match their students' linguistic level. Such adaptation results in inconsistent assessment standards across school types. In short, the national policy goal of standardization in speaking assessment remains difficult to realize under current structural conditions. Addressing these differences would require differentiated support, including more instructional time for English, adjusting the level of curriculum according to the differences among school types, accessible rater-training resources, and additional materials like authentic tasks and detailed rubrics that promote comparable testing practices across institutions.

In conclusion, this study has shown that the validity and reliability of classroom-based English-speaking assessments are shaped as much by contextual realities as by theoretical principles. While teachers generally succeed in creating exams that are familiar, age-appropriate, and aligned with curricular goals, deeper dimensions of validity, particularly construct coverage and scoring consistency, remained limited. The findings reveal that time constraints, possible mismatch between students and the curriculum, lack of standardized

rubrics, and the absence of moderation procedures continue to hinder the development of robust assessment practices across school types. Yet, the moderate levels of agreement among experts regarding the face and content validity constructs suggest a strong foundation upon which improvement can be built. Ultimately, by bridging the gap between theory and practice, schools can move toward assessment systems that not only measure speaking ability more accurately but also foster learners' confidence and communicative competence, the true aim of language education.

5.4. Limitations of the study

Despite the informative nature of this study, several limitations must be acknowledged. First, the study relied on the evaluations of four expert raters.

Second, the assessment of the speaking exams was based on documentation (e.g., test papers, rubrics, etc.), not on live or recorded student performances. As a result, certain dynamic aspects of speaking assessments, such as test-taker anxiety, teacher behavior, or classroom conditions, could not be evaluated.

Third, the "I don't know" responses, while methodologically excluded from Kappa calculations, may still reflect ambiguity in item wording or lack of information provided by schools.

The study does not provide evidence regarding response or washback validity, as no data were collected on how test-takers interpreted or approached the test tasks.

The study did not include follow-up questions which may have been used during the tests in the analysis, because of their inherent variability along with the impracticality of collecting such data.

Finally, the study focused exclusively on high school English speaking exams in Sakarya province of Türkiye within the Turkish education context.

5.5. Implications for practice

The findings of this study carry several practical implications for educators, test developers, and policymakers seeking to improve the quality and fairness of school-based speaking assessments.

First, results showed that teachers often simplified tasks, reduced linguistic demands, or provided highly guided prompts in response to perceived student proficiency levels, which contributed to low construct validity and score inconsistencies. This highlights the need for sustained professional development in assessment literacy. Teachers who design and administer speaking exams should receive targeted training addressing key psychometric principles such as construct validity, rater reliability, and rubric use (Giraldo, 2021). Short, practice-oriented workshops supported by the Ministry of National Education (MoNE) could help teachers move beyond intuitive scoring and adopt more systematic, evidence-based procedures. Integrating language testing theory and practice into pre-service teacher education programs would further strengthen this foundation.

Second, expert evaluations indicated misalignment between rubric descriptors, intended learning outcomes, and curriculum expectations, resulting in inconsistent scoring across schools. The study shows the importance of standardization and detailed rubric development. Comprehensive analytic rubrics aligned with CEFR levels and curriculum outcomes should be provided centrally and their consistent use encouraged across schools, enhancing both transparency and inter-rater reliability (Limgomolvilas and Sukserm 2025). Additionally, the MoNE could introduce quality assurance frameworks to ensure alignment between teacher-made test design practices and national language education policies.

Third, findings revealed that none of the schools implemented reliability-enhancing procedures such as test-retest, parallel (equivalent) forms, split-half methods, multiple raters, repeated scoring, or external raters. This suggests that institutional monitoring and collaborative moderation mechanisms could play a crucial role in maintaining assessment quality and fairness. Establishing moderation systems or peer-review processes in which teachers jointly evaluate sample performances can promote a shared understanding of assessment standards.

5.6. Suggestions for future research

Findings revealed that monologic and memorization based tasks, frequently used in several schools, were associated with lower construct validity. Future research should experimentally compare alternative task formats, such as performance-based, scenario-driven, or interactive speaking tasks, to examine how these formats influence the range of competencies elicited. Controlled quasi-experimental designs or mixed-method studies incorporating live or video-recorded assessments could provide insight into how task type affects student performance and the representativeness of the assessed construct.

The current study highlighted misalignments between the intended learning outcomes and rubric descriptors, particularly in relation to national curriculum and CEFR. Future studies should focus on the development and empirical validation of rubrics that systematically operationalize curriculum and CEFR descriptors. Pilot studies or multi-site experimental designs could assess whether these tools more accurately capture intended competencies and enhance the consistency of scoring.

Expert evaluations indicated that teachers' tendencies to simplify tasks, reduce linguistic demands, or provide highly guided prompts affected cognitive difficulty and construct representation. These practices, while intended to support students, limited the scope of assessed skills and contributed to score inconsistencies. Future research should employ mixed-methods designs combining classroom observation, think-aloud protocols, and teacher interviews to investigate how task design, teacher behavior, and school-type differences jointly impact construct validity and reliability in speaking assessment. Comparative studies across academically selective and general schools could illuminate systemic patterns influencing assessment quality. In other words, future studies may investigate how teachers negotiate curriculum expectations, student needs, and assessment pressures, and how these negotiations influence the validity of the evaluation process.

Given the absence of live performance data in the current study, future research should incorporate observational or video-based datasets. Smaller, controlled sample studies could allow for detailed analyses of student-test task interactions, examiner interventions, and performance features that cannot be captured through document analysis alone. Longitudinal or multi-site designs would enable researchers to test whether improvements in assessment tools and instructional practices translate into more consistent validity and reliability outcomes across school types.

A key finding of the study was that none of the schools employed established reliability-enhancing methods such as test-retest, parallel (equivalent) forms, split-half procedures, multiple raters, repeated scoring, or external raters. Future research should systematically investigate the feasibility and impact of incorporating these procedures in high school speaking assessments. Experimental or quasi-experimental studies could compare the effects of different reliability strategies on inter-rater consistency and score stability across diverse school contexts. Additionally, studies could explore practical guidelines for implementing these methods within the constraints of regular classroom schedules and teacher workloads, providing evidence-based recommendations for improving both reliability and construct validity in speaking assessments.

REFERENCES

Ahmadi, A., & Sadeghi, E. (2016). Assessing English language learners' oral performance: A comparison of monologue, interview, and group oral test. *Language Assessment Quarterly, 13*(4), 341-358.

Airasian, P. ve Russel, M. (2012). *Classroom assesment: concepts and applications*. (7th ed.). Boston, MA: McGraw-Hill Higher Education.

Ak, M., & Bozyiğit, E. (2024). Decoding teacher communities. *AELTE 2024*, 9.

Akçay, A., Tunagür, M., & Karabulut, A. (2020). Turkish teachers' assessment situations: A study on exam papers. *International Journal of Education and Literacy Studies, 8*(2), 36-43. doi:https://doi.org/10.7575/aiac.ijels.v.8n.2p.36

Akhter, S. (2021). Exploring the significance of speaking skill for EFL learners. *SJESR, 4*(3), 1–9. DOI**:** https://doi.org/10.36902/sjesr-vol4-iss3-2021(1-9)

Akıncı, B. (2020). *Fen Bilimleri Dersi Öğretim Programı ve Ölçme Değerlendirme Araçlarının Akademik Becerilerin İzlenmesi ve Değerlendirilmesine (Abide) Göre İncelenmesi*. (Master's thesis) Retrieved from YÖK Thesis Center (Thesis No:617651).

Alderson, J. C. (2000). *Assessing reading.* Cambridge, UK: Cambridge University Press.

Alderson, C., & Wall, D. (1993). Does washback Exist? *Applied Linguistics, 14* (2), 115-129. https://doi.org/10.1093/applin/14.2.115

Alderson, J.C., Clapham, C. & Wall, D. (1995). *Language test construction and evaluation*. Cambridge, UK: Cambridge University Press.

Aliqi, G., & Hayes, E. (2024). *Task-based learning and oral communication skills* (Master's thesis, Malmö University) Retrieved from: http://www.diva-portal.org/smash/get/diva2:1929545/FULLTEXT02.pdf

Allan, D. (1999). Testing and assessment. *English Teaching Professional*, 11, 19-20.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Anderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.

Arslan, R. Ş., & Üçok Atasoy, M. (2020). An investigation into EFL teachers' assessment of young learners of English: Does practice match the policy? *International Online Journal of Education and Teaching (IOJET). 7(2), 468-484. http://iojet.org/index.php/IOJET/article/view/818*

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford University Press.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, UK: Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford University Press.

Bademci, V. (2019). Geçerlik: Nedir? Ne Değildir? *Eğitim ve Toplum Araştırmaları Dergisi, 6*(2), 373–385. https://dergipark.org.tr/en/pub/etad/issue/51092/646773

Ball, S. J., Maguire, M., & Braun, A. (2012). *How schools do policy: Policy enactments in secondary schools.* London: Routledge.

Ballıdağ, S., & İnan Karagül, B. (2021). Exploring the language assessment literacy of Turkish in-service EFL teachers. *Balıkesir Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, *24*(45), 73–92. https://doi.org/10.31795/baunsobed.909953

Baykul, Y. (2000). *Eğitimde ve Psikolojide Ölçme: Klasik Test Teorisi ve Uygulaması* (1.Baskı). Ankara: Pegem Akademi.

Baykul, Y. (2010). *Eğitimde ve Psikolojide Ölçme: Klasik Test Teorisi ve Uygulaması* (2.Baskı). Ankara: Pegem Akademi.

Bejar, I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice, 3*1(3), 2-9.

Berry, V. (2007). *Personality differences and oral test performance*. Frankfurt: Peter Lang.

Board of Education - Ministry of National Education (Milli Eğitim Bakanlığı Talim ve Terbiye Kurulu Başkanlığı) *Haftalık ders çizelgeleri*. (20.05.2025) https://ttkb.meb.gov.tr/www/haftalik-ders-cizelgeleri/kategori/7 [Accessed 17 May 2025].

Bowen, G. A. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal, 9*(2), 27–40. https://doi.org/10.3316/QRJ0902027

Brindley, G. (1991). Defining language ability: the criteria for criteria. In S. Anivan (Ed.), *Current developments in language testing (139-164).* Singapore: Regional Language Centre. Retrieved from: https://eric.ed.gov/?id=ED365150

British Educational Research Association (BERA). (2018). *Ethical guidelines for educational research* (4th ed.). BERA.

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. USA-ASDC.

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: co-constructing a better performance. *Language Testing*, 26(3), 341-366. http://dx.doi.org/10.1177/0265532209104666.

Brown, A. (2000). An investigation of the rating process in the IELTS oral interview. *International English Language Testing System (IELTS) Research Reports, 3*(1), 49.

Brown, A. (2005). *Interviewer variability in oral proficiency interviews*. Frankfurt: Peter Lang.

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, 2005*(1), i-157. Retrieved from: https://files.eric.ed.gov/fulltext/EJ1111373.pdf

Brown, A., McNamara, T., Iwashita, N. and O'Hagan, S. (2001). Investigating raters' orientations in specific-purpose task-based oral assessment. (TOEFL: research and development project reports). Princeton, NJ: Educational Testing Service.

Brown, H.D. (2004). *Language assessment. Principles and classroom practices*. NY, USA: Pearson Education.

Brown, H. D., & Abeywickrama, P. (2010). *Language assessment: Principles and classroom practices* (2nd ed.). Pearson Longman.

Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice-Hall.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge University Press.

Brown, J. D. & Baile, K. M. (2008). Language testing courses: What are they in 2007. *Language Testing*, 25(3), 349–383. https://doi.org/10.1177/0265532208090157

Buck, G. (2001). *Assessing listening.* Cambridge, UK: Cambridge University Press.

Büyüköztürk, Ş., Kılıç Çakmak, E., Akgün, Ö. E., Karadeniz, Ş., & Demirel, F. (2011). *Eğitimde bilimsel araştırma yöntemleri*. Pegem Akademi.

Bygate, M. (Ed.) (1987). *Speaking.* Oxford: Oxford University Press.

Campbell, E. H. (1993). *Fifteen Raters Rating: An Analysis of Selected Conversation during Placement Rating Session*. Speech presented at The Annual Meeting of the Conference on College Composition and Communication, San Diago, CA.

Carey, L. M. (1988). *Measuring and evaluating school learning*. Newton, Massachusetts: Allyn and Bacon, Inc

Cesur, Ç. (2009). An assessment of the validity of the standardized achievement test administered at Çanakkale Onsekiz Mart University (Master's thesis) Retrieved from YÖK Thesis Center (Thesis No:249197).

Chalhoub-Deville, M., & Wigglesworth, G. (2005). Rater judgment and English language speaking proficiency. *World Englishes, 24*(3), 383-391. https://doi.org/10.1111/j.0083-2919.2005.00419.x

Cheng, L. & Curtis, A. (2004). Washback or backwash: a review of the impact of testing on teaching and learning. In L.

Cheng & Y. Watanabe (with A. Curtis) (Eds.), *Washback in language testing: research contexts and methods* (3-17). Mahwah, New Jersey: Lawrence Erlbaum.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Cizek, G. J. (2010). An introduction to formative assessment: History, characteristics, and challenges. In *Handbook of formative assessment* (pp. 3-17). Routledge.

Clark, J. L. D. (1978). *Direct Testing of Speaking Proficiency: Theory and Application*. Princeton, NJ: Educational Testing Service.

Cohen, A. D. 1994 *Assessing Language Ability in the Classroom*. BOSTON: Heinle & Heinle.

Cohen, L., Manion, L., & Morrison, K. (2002). *Research methods in education*. Routledge.

Collie, J. (1998). *Double Take 2: Language Practice: Skills Training*. Oxford University Press.

Coombe, C.; Troudi, S. & Al-Hamly, M. (2012). Foreign and second language teacher assessment literacy: issues, challenges and recommendations. In Coombe, C, Davidson, P, O'Sullivan, B, and Stoynoff, S (eds.) *The Cambridge guide to second language assessment*. Cambridge: Cambridge University Press.

Creswell, J. W. (2013). *Qualitative inquiry and research design: Choosing among five approaches* (3rd ed.). Thousand Oaks, CA: Sage.

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE Publications.

Crocker, L. ve Algina, J. (1986). *Introduction of classical and modern test theory*. Ohio: Cengage Learning.

Crystal, D. (2003). *English as a global language* (2nd ed.). Cambridge University Press.

Çaykan, I. (2001). Material development for testing oral skills communicatively for intermediate students. (Unpublished master's thesis). Hacettepe University, Ankara, Turkey.

Çimen, Ş. S. (2022). Exploring EFL assessment in Turkey: Curriculum and teacher practices. *International Online Journal of Education and Teaching (IOJET), 9*(1), 531–550 Retrieved from: https://files.eric.ed.gov/fulltext/EJ1327768.pdf

Congdon, P.J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement*, 37(2), 163–178.

Çopur, D. (2002). Testing first year FLE students' oral performance using four speaking test methods in spoken English II course and students' attitudes towards these speaking testing methods (Unpublished MA thesis). Middle East Technical University, Ankara.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. and McNamara, T. (1999). *Dictionary of language testing*. Cambridge: Cambridge University Press.

Davis, L. E. (2012). *Rater expertise in a second language speaking assessment: The influence of training and experience* (Unpublished doctoral dissertation). University of Hawai'i at Manoa, Hawaii.

Doğru, Ş.C. (2019). *Karma Testlerin Psikometrik Özelliklerini Belirlemede Klasik Test Kuramı ve Rasch Modelinin Karşılaştırılması* (Master's thesis) Retrieved from YÖK Thesis Center (Thesis No:586589).

Douglas, D. (Eds.) (2014). *Understanding language testing (Second Edition).* New York, USA: Routledge.

Durai, D. N., & Soundrarajan, A. (2009). Importance of English language. *International Journal of Research in Humanities, Arts and Science*, 38.

Dursun, F., Bedir, S. B., & Gülcü, E. Ö. (2017). LİSE İNGİLİZCE DERSİ ÖĞRETMENLERİNİN ÖĞRETİM PROGRAMINA İLİŞKİN GÖRÜŞLERİNİN İNCELENMESİ. *Milli Eğitim Dergisi*, 46(216), 135-163.

Dutta, S. (2019). The importance of "English" language in today's world. *International Journal of English Learning & Teaching Skills*, *2*(1), 1028-1035 DOI:10.15864/ijelts.2119

East, M. (2016). Assessing foreign language students' spoken proficiency: stakeholder perspectives on assessment innovation. Singapore: Springer.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Englewood Cliffs, NJ: Prentice Hall.

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly*, *9*(3), 270-292. https://doi.org/10.1080/15434303.2011.649381

Ekmekçi, E. (2016). Comparison of native and non-native English language teachers' evaluation of EFL learners' speaking skills: Conflicting or identical rating behaviour? *English Language Teaching*, *9*(5), 98–105. DOI:10.5539/elt.v9n5p98

Engelhard, G. Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*(2), 93-112. https://doi.org/10.1111/j.1745-3984.1994.tb00436.x

Field, A. (2018). *Discovering statistics using IBM SPSS Statistics* (5th ed.). Sage Publications.

Fitzpatrick, R. & Morrison, E.J. (1971). Performance and product evaluation. In R. L. Thorndike (Ed.), Educational measurement (pp. 237-270). Washington DC: American Council of Education.

Fulcher, G. (2003). *Testing Second Language Speaking*. London: Pearson Education Limited.

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education* (8th ed.). McGraw-Hill.

Friedman, T. L. (2005). *The world is flat: A brief history of the twenty-first century*. Farrar, Straus and Giroux.

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

Fulcher, G. (Ed.) (2014). Testing second language speaking (Second Edition). London & New York: Routledge.

Fulcher, G. & Davidson, F. (2007). Language testing and assessment: an advanced resource book. New York: Routledge.

Galaczi, E. D. (2008). Peer–peer interaction in a speaking test: The case of the First Certificate in English examination. *Language Assessment Quarterly, 5*(2), 89–119. https://doi.org/10.1080/15434300801934702

Gampper, C. (2013). Improving English test qualities. *Thammasat Review, Special Issue,* 73-83

Gareis, C., & Grant, L. W. (2008). *Teacher-made assessments: How to connect curriculum, instruction, and student learning*. Eye On Education.

Genç, E., Çalışkan, H., & Yüksel, D. (2020). Language assessment literacy level of EFL teachers: A focus on writing and speaking assessment knowledge of the teachers. *Sakarya University Journal of Education, 10*(2), 274-291. https://doi.org/10.19126/suje.626156

Getman, E. P. (2020). Age, task characteristics, and acoustic indicators of engagement: Investigations into the validity of a technology-enhanced speaking test for young language learners (Doctoral dissertation, Teachers College, Columbia University).

Ginty, A. T. (2020). Psychometric properties. In M. D. Gellman (Ed.), *Encyclopedia of behavioral medicine*. Springer. https://doi.org/10.1007/978-3-030-39903-0_480

Giraldo, F. (2021). Language assessment literacy and teachers' professional development: A review of the literature. *Profile Issues in Teachers Professional Development, 23*(2), 265-279. DOI: https://doi.org/10.15446/profile.v23n2.90533

Gökdemir, C. V. (2010). Üniversitelerimizde verilen yabancı dil öğretimindeki başarı durumumuz. *Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 6*(2), 251–264. Retrieved from: https://dergipark.org.tr/en/pub/ataunisosbil/issue/2816/37912

Graddol, D. (2006). *English next: Why global English may mean the end of 'English as a foreign language'*. British Council.

Green, A. (2014). *Exploring language assessment and testing: Language in action*. London: Routledge.

Guilford, J. P. (1954). *Psychometric methods* (2d Ed). New York: McGraw-Hill.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*(1), 17–27. https://doi.org/10.1111/j.1745-3992.2004.tb00149.x

Han, Q. (2016). Rater cognition in l2 speaking assessment: A review of the literature. Teachers College, Columbia University *Working Papers in TESOL & Applied Linguistics*, *16*(1), 1-24.

Harmer, J. (2004). *The practice of English language teaching*. Longman.

Hasselgreen, A.; Carlsen, C. & Helness, H. (2004). *European survey of language testing and assessment needs*. Report: Part one, general findings, University of Bergen

Hatipoğlu, Ç. (2015). English language testing and evaluation (ELTE) training in Turkey: expectations and needs of pre-service English language teachers. *International Association of Research in Foreign Language Education and Applied Linguistics ELT Research Journal*, 4(2), 111-128 Retrieved from: https://dergipark.org.tr/en/pub/eltrj/issue/28780/308006

Haznedar, B. (2010). Türkiye'de yabancı dil eğitimi: reformlar, yönelimler ve öğretmenlerimiz. *International Conference on New Trends in Education and their Implications 11-13 November*, Antalya, Turkey

Heaton, J.B. (Eds.) (1988). *Writing English language tests (Second Edition).* New York: Longman.

Henning, G. (1987). *A Guide to language testing: development, evaluation and research*. Rowley, Massachusetts: Newbury House.

Herrera Mosquera, L., & Macías V, D. F. (2015). A call for language assessment literacy in the education and development of teachers of English as a foreign language. *Colombian Applied Linguistics Journal, 17*(2), 302–312.https://doi.org/10.14483/udistrital.jour.calj.2015.2.a09

Höl, D. (2010). Perceptions and attitudes of the English language instructors and preparatory students towards testing speaking communicatively in the school of foreign languages at Pamukkale University. (Unpublished master's thesis). Pamukkale University, Denizli, Turkey.

Höl, D. (2018). Current problems and proposed solutions for testing speaking: Opinions of EFL teachers. *The Literacy Trek, 4*(1), 27–36. Retrieved from: https://dergipark.org.tr/en/pub/literacytrek/issue/38066/350216

Huang, H. T. D., Hung, S. T. A., & Hong, H. T. V. (2016). Test-taker characteristics and integrated speaking test performance: A path-analytic study. *Language Assessment Quarterly, 13*(4), 283–301. https://doi.org/10.1080/15434303.2016.1236111

Hubackova, S. (2016). The importance of foreign language education. *New Trends and Issues Proceedings on Humanities and Social Sciences, 2*(5). Retrieved from https://un-pub.eu/ojs/index.php/pntsbs/article/view/1108 (Original work published January 12, 2017)

Huerta-Macias, A. (1995). Alternative assessment: responses to commonly asked questions. *TESOL Journal*, 5(1), 8-11. DOI:10.1017/CBO9780511667190.048

Hughes, A. (2003). *Testing for language teachers*. Cambridge University Press.

Işık, A. (2005). Egemen dilin diğer dilleri etkisi altına alması ve yabancı dil eğitimi. *İzzet Baysal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 11*, 83–102. https://doi.org/10.11616/AbantSbe.144

Işık, A. (2008). Yabancı dil eğitimimizdeki yanlışlar nereden kaynaklanıyor? *Journal of Language and Linguistic Studies, 4*(2), 15–26. Retrieved from: https://dergipark.org.tr/tr/pub/jlls/issue/9929/122850

Işık, E., & Semerci, Ç. (2019). Nitel araştırmalarda veri üçgenlemesi olarak odak grup görüşmesi, bireysel görüşme ve gözlem. *Turkish Journal of Educational Studies, 6*(3), 53–66. https://doi.org/10.33907/turkjes.607997

İleri, S. A. (2019). İngilizce öğretmenlerinin konuşma becerisi ölçümüne ilişkin gereksinimleri ve bir hizmet içi eğitim önerisi (Doctoral dissertation) Retrieved from YÖK Thesis Center (Thesis No:588085).

İnal, E. (2020). Yabancı dil öğretiminde konuşma becerisinin geliştirilmesi ve ölçme-değerlendirme süreçlerine yönelik düşünceler. *Aydın TÖMER Dil Dergisi, 5*(2), 189–204. Retrieved from: https://dergipark.org.tr/tr/pub/aydintdd/issue/60730/851465

İnan, T. (2019). Rater cognition in the assessment of EFL speaking performance (Master's thesis) Retrieved from YÖK Thesis Center (Thesis No:571019).

Jonsson, A. & Svingby, G. (2007). The use of scoring rubrics: reliability, validity and educational consequences. *Educational Research Review-2*, 130-144 https://doi.org/10.1016/j.edurev.2007.05.002

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Praeger.

Kaplan, R.M. & Saccuzzo, D.P. (2001). *Psychological testing: Principle, applications and issues* (5th Ed.). Belmont, CA: Wadsworth.

Karakaş, A. (2012). Evaluation of the English language teacher education program in Turkey. *ELT Weekly, 4*(15), 1-16. Retrieved from: https://www.researchgate.net/publication/272493636_Evaluation_of_the_English_Language_Teacher_Education_Program_in_Turkey

Kılıç, A.F. (2015). *Temel Eğitimden Ortaöğretime Geçiş Ortak ve Mazeret Sınavındaki Türkçe ve Matematik Alt Testlerinin Psikometrik Özelliklerinin Karşılaştırılması* (Yüksek Lisans Tezi). (Master's thesis) Retrieved from YÖK Thesis Center (Thesis No:418186).

Kıral, B. (2020). Nitel bir veri analizi yöntemi olarak doküman analizi. *Siirt Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 8*(15), 170–189. Retrieved from: https://dergipark.org.tr/tr/pub/susbid/issue/54983/727462

Kızıldağ, A. (2009). Teaching English in Turkey: Dialogues with teachers about the challenges in public primary schools. *International Electronic Journal of Elementary*

Education, *1*(3), 188–201. Retrieved from https://www.iejee.com/index.php/IEJEE/article/view/274

Kim, H. J. (2015). A qualitative analysis of rater behavior on an L2 speaking assessment. *Language Assessment Quarterly*, *12*(3), 239-261. DOI:10.1080/15434303.2015.1049353

Kitao, S.K. & Kitao, K. (1998). Testing communicative competence. The Internet TESL Journal, 2, 1-4. Retrieved from: https://files.eric.ed.gov/fulltext/ED398260.pdf

Koizumi, R. (2022). L2 speaking assessment in secondary school classrooms in Japan. *Language Assessment Quarterly*, *19*(2), 142–161. https://doi.org/10.1080/15434303.2021.2023542

Köksal, D. (2004). Assesing teachers' testing skills in elt and enhancing their professional development through distance learning on the net. *Turkish Online Journal of Distance Education-TOJDE* *5*(1) Retrieved from: https://files.eric.ed.gov/fulltext/ED494459.pdf

Kremmel, B., & Harding, L. (2020). Towards a theory of assessment literacy: Learning from language testing and assessment. *Language Testing*, 37(3), 396–416.

Kyle, K., Crossley, S. A., & McNamara, D. S. (2016). Construct validity in TOEFL iBT speaking tasks: Insights from natural language processing. *Language Testing*, *33*(3), 319–340. https://doi.org/10.1177/0265532215587391

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Lane, S., & Stone, C.A. (2006). Performance Assessment. In R. L. Brennan (Ed.): *Educational Measurement* (pp 387-431). Wesport, CT: ACE/Praeger.

Limgomolvilas, S. & Sukserm, P. (2025). Examining rater reliability when using an analytical rubric for oral presentation assessments. *LEARN Journal: Language Education and Acquisition Research Network*, *18*(1), 110-134. https://doi.org/10.70730/JQGY9980

Luoma, S. (2004). *Assessing speaking*. Cambridge University Press.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to raters? *Language Testing 19/3:* 246-276. https://doi.org/10.1191/0265532202lt230oa

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12/1*: 54–71. https://doi.org/10.1177/026553229501200104

Madsen, H.S. (1983). *Techniques in testing.* Oxford: Oxford University Press.

Majidifard, E., Shomoossi, N. & Ghourchaei, B. (2014). Risk taking, gender and oral narrative proficiency in Persian. *Procedia - Social and Behavioral Sciences*, 98, 1085-1092. Retrieved from: https://www.sciencedirect.com/science/article/pii/S1877042814026111/pdf?md5=8991e07a6998412ea904a40bd346a25c&pid=1-s2.0-S1877042814026111-main.pdf

May, L. (2009). Co-constructed interaction in a paired speaking test: the rater's perspective. Language Testing, 26(3), 397–422. http://dx.doi.org/10.1177/0265532209104668.

McEwen, N. (1995). Educational accountability in Alberta. *Canadian Journal of Education*, 20, 27-44. Retrieved from: https://journals.sfu.ca/cje/index.php/cje-rce/article/view/2702

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

McNamara, T. & Shohamy, E. (2008). Language tests and human rights. *International Journal of Applied Linguistics*, 18(1), 89-95. https://doi.org/10.1111/j.1473-4192.2008.00191.x

MEB (2023). MİLLÎ EĞİTİM BAKANLIĞI YAZILI VE UYGULAMALI SINAVLAR YÖNERGESİ (Directive for Written and Practical Exams) https://odsgm.meb.gov.tr/meb_iys_dosyalar/2023_10/12115933_MEB_yazili_ve_uygulamali_sinavlar_yonergesi.pdf [Accessed 10 March 2025].

MEB (2024) 2021-2023 Yıllarında Planlanan Hizmet içi Eğitim Faliyetleri Bilgileri Retrieved from: https://oygm.meb.gov.tr/www/hizmetici-egitim-planlari/icerik/28

Mede, E., & Atay, D. (2017). English language teachers' assessment literacy: The Turkish context. *Dil Dergisi*, *168*(1), 1-5. Retrieved from: https://dergipark.org.tr/en/download/article-file/780021

Mehnert, U. (1998). The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition*, 20, 83-108. Retrieved from: http://www.jstor.org/stable/44486384

Merriam, S. B. (1998). *Qualitative research and case study applications in education*. San Francisco, CA: Jossey-Bass.

Morrow, K. (2012). Communicative language testing. In C. Coombe, P. Davidson, B. O'Sullivan and S. Stoynoff (Eds), *The Cambridge guide to second language assessment (140-146*). Cambridge: Cambridge University Press.

Moskal, B. M. (2000). Scoring rubrics: What, when and how? *Practical Assessment, Research & Evaluation, 7*(3). Retrieved from: http://PAREonline.net/getvn.asp?v=7&n=3

Mousavi, S.A. (2002). *An encyclopedic dictionary of language testing* (Third Edition). Taiwan: Tung Hua Book Company.

Murphy, K. R. & Davidshofer, C. O. (2004). *Psychological testing principles and applications* (6. Edition). Phoenix: Prentice Hall.

Myford, C. M. (2012). Rater cognition research: Some possible directions for the future. *Educational Measurement: Issues and Practice, 31*(3), 48-49. https://doi.org/10.1111/j.1745-3992.2012.00243.x

Myford, C.M., & Wolfe, E.W. (2004). *Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part I*. In E. V. Smith y R.M. Smith (Eds.). Introduction to Rasch Measurement (pp. 460-515). Maple Grove, MN: JAM Press.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*(2), 189-227. Retrieved from: https://www.researchgate.net/publication/9069043_Detecting_and_Measuring_Rater_Effects_Using_Many-Facet_Rasch_Measurement_Part_I

Norris, J. M., Brown, J. D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments*. University of Hawai'i, Second Language Teaching & Curriculum Center

Ohiri, A. C., & Ihebom, A. P. (2024). *Psychometric properties of a test: An overview*.

O'Malley, M. & Valdez-Pierce, L. (1996). *Authentic assessment for English language learners: practical approaches for teachers*. New York: Addison Wesley.

O'Sullivan, B. (2006). *Modelling performance in oral language tests: language testing and evaluation*. Frankfurt: Peter Lang.

O'Sullivan, B. (2013). Assessing spoken language: Scoring validity. In D. Tsagari, S. Papadima-Sophocleous, S. Ioannou-Georgiou (Eds.), *Language Testing and Evaluation: international experiences in language testing and assessment* (Volume 28, pp. 261-274). Frankfurt: Peter Lang Edition.

O'Sullivan, B. and Weir, C. (2002). Research issues in testing spoken language. (Unpublished research report). Cambridge: University of Cambridge Local Examinations Syndicate.

Ösken, H. (1999). An assessment of the validity of the midterm and the end of course assessment tests administered at Hacettepe University, Department of Basic English (Master's thesis) Retrieved from YÖK Thesis Center (Thesis No:89709).

Öz, H. (2014). Turkish teachers' practices of assessment for learning in the English as a foreign language classroom. *Journal of Language Teaching & Research*, *5*(4), 775-785. Retrieved from: https://www.academypublication.com/issues/past/jltr/vol05/04/07.pdf

Özdemir, C. (2018). Assessing the speaking skill: An investigation into achievement tests of the 9th grade students in Anatolian high schools (Master's Thesis) Retrieved from YÖK Thesis Center (Thesis No:525526).

Paker, T. (2007). Problems of teaching English in schools in Çal Region and suggested solutions. In *21. yüzyıla girerken geçmişten günümüze Çal yöresi: Baklan, Çal, Bekilli* (Vol. 3, pp. 684–690). Çal Yöresi Yardımlaşma ve Dayanışma Derneği Yayını. Retrieved from: https://www.researchgate.net/publication/258028815_Bridging_the_Gap_between_Policy_and_Practice_in_Teaching_English_to_Young_Learners_the_Turkish_Context

Pineda, D. (2014). The feasibility of assessing teenagers' oral english language performance with a rubric. *Profile Issues in Teachers' Professional Development, 16*(1), 181-198. Retrieved from: https://eric.ed.gov/?id=EJ1053772

Polat, M. (2020). A Rasch analysis of rater behaviour in speaking assessment. *International Online Journal of Education and Teaching (IOJET), 7*(3), 1126–1141. Retrieved from: https://search.trdizin.gov.tr/tr/yayin/detay/376514

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health, 29*(5), 489–497. DOI: 10.1002/nur.20147

Rajan, N. (2013, November). Turkey and BRICs gain in latest EF English Proficiency Index Retrieved from: https://thepienews.com/turkey-and-brics-gain-in-latest-ef-english-proficiency-index/

Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education, 35*(4), 435-448. https://doi.org/10.1080/02602930902862859

Regulation for Secondary Education Institutions [Ortaöğretim Kurumları Yönetmeliği]. (2016). T.C. Resmi Gazete [T.R. Official Gazette], 29871, 28 October 2016. https://ogm.meb.gov.tr/meb_iys_dosyalar/2016_11/03111224_ooky.pdf [Accessed 9 February 2025].

Restrepo, A.P.M., Aristizábal, L.D., Orozco, F.C., Monsalve, S.G., Orozco, L.A.L. and Urán, M.P. (2003). Assessing spoken language in EFL: beliefs and practices. *Revista Universidad EAFIT*, [S.l.], 39(129), 63-73. Retrieved from: https://www.redalyc.org/pdf/215/21512906.pdf

Roca-Varela, M. L. & Palacios, M.I. (2013). How are spoken skills assessed in proficiency tests of general English as a foreign language? A preliminary survey. *International Journal of English Studies, 13*(2),53-68. Retrieved from: https://eric.ed.gov/?id=EJ1042855

Sak, G. (2008). An investigation of the validity and reliability of the speaking exam at a Turkish university. (Unpublished master's thesis). Middle East Technical University, School of Social Sciences, Ankara, Turkey.

Sarıçoban, A. (2011). A study on the English language teachers' preparation of tests. *Hacettepe University Journal of Education*, 41, 398-410. Retrieved from: http://efdergi.hacettepe.edu.tr/yonetim/icerik/makaleler/709-published.pdf

Semiz, Ö., & Odabaş, K. (2016). Turkish EFL teachers' familiarity with and perceived needs for language testing and assessment literacy. In *Proceedings of the Third International Linguistics and Language Studies Conference* (pp. 66–72). Retrieved from: https://www.researchgate.net/profile/Iga-Lehman/publication/337482422_The_influence_of_writing_on_thinking_patterns_and_expression_in_a_cross-cultural_perspective/links/5e17696ca6fdcc283763a634/The-influence-of-writing-on-thinking-patterns-and-expression-in-a-cross-cultural-perspective.pdf#page=66

Sheng-ping, T. & Chong-ning, X. (2004, May). On washback of testing to general English education. The Fourth International Conference in ELT, China.

Shohamy, E. (1983). Interrater and intrarater reliability of the oral interview and concurrent validity with cloze procedure in Hebrew. In J.W.Oller (ed.). Issues in Language Testing Research. Rowley, MA: Newbury House.

Shohamy, E. (1988). A proposed framework for testing the oral language of second/foreign language learners. *Studies in Second Language Acquisition*, 10, 165-179.

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing*, 11, 99-123.

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. Longman.

Shohamy, E., Donitsa-Schmidt, S. & Ferman, I. (1996). Test impact revisited: washback effect over time. *Language Testing*, 13, 298–317. https://doi.org/10.1177/026553229601300305

Skehan, P. (1996). A framework for the implementation of task based instruction. *Applied Linguistics*, 17, 38-62. DOI:10.1093/applin/17.1.38

Skehan, P. and Foster, P. (1997). The influence of planning and post-task activities on accuracy and complexity in task-based learning. Language Teaching Research, 1(3), 185-211. DOI:10.1191/136216899672186140

Sook, K.H. (2003). Korean junior secondary school English teachers' perceptions of speaking assessment. *Asian EFL Journal*, 5(4), 1-30. DOI:10.1186/2229-0443-2-4-75

Spolsky, B. (2008). Language testing at 25: maturity and responsibility. *Language Testing*, 25(3), 297–305. https://doi.org/10.1177/0265532208090153

Sreena, S., & Ilankumaran, M. (2018). Developing productive skills through receptive skills–a cognitive approach. *International Journal of Engineering & Technology*, *7*(4.36), 669-673. Retrieved from: https://www.academia.edu/download/103667327/12213.pdf

Stevens, D. D., & Levi, A. J. (2004). *Introduction to rubrics: an assessment tool to save grading time, convey effective feedback, and promote student learning*. Sterling, VA: Stylus Publishing.

Stiggins, R. & Faires-Conkin, N. (1992). *In teachers' hands*. Albany, NY: State University of New York Press.

Şallı-Çopur, D. (2002). Testing first year FLE students' oral performance using four speaking test methods in spoken English II course and students' attitudes towards these speaking test methods.

Tavares, W., & Eva, K. W. (2013). Exploring the impact of mental workload on raterbased assessments. *Advances in Health Sciences Education, 18*(2), 291-303. DOI: 10.1007/s10459-012-9370-3

Teddlie, C., & Yu, F. (2007). Mixed methods sampling: A typology with examples. *Journal of Mixed Methods Research, 1*(1), 77–100. DOI 10.1177/1558689806292430

Tekin, H. (1996). *Eğitimde ölçme ve değerlendirme*. Yargı Yayınları.

Tılfarlıoğlu, F., & Öztürk, A. (2007). An analysis of ELT teachers' perceptions of problems concerning the implementation of English language teaching curricula in elementary schools. *Journal of Language and Linguistic Studies, 3*(1), 202–217. Retrieved from: https://dergipark.org.tr/en/pub/jlls/issue/9925/122874

Turgut, F. (1990). *Eğitimde ölçme değerlendirme metotları*. Yargıcı Matbaası.

Türnüklü, A. (2000). Eğitimbilim araştırmalarında etkin olarak kullanılabilecek nitel bir araştırma tekniği: Görüşme. *Kuram ve Uygulamada Eğitim Yönetimi, 24*(24), 543–559. Retrieved from: https://dergipark.org.tr/tr/pub/kuey/issue/10372/126941

Underhill, N. (1987). *Testing spoken language: a handbook of oral testing techniques*. Cambridge: Cambridge University Press.

UNESCO, (2024) Literacy rate, adult total (% of people ages 15 and above) Retrieved from: https://data.worldbank.org/indicator/SE.ADT.LITR.ZS [Accessed 10 April 2025].

Uzun, A., & Kılıçkaya, F. (2020). TEOG (TEPSE) English test: Content validity and teachers' views. *Bartın University Journal of Faculty of Education, 9*(3), 680–708. https://doi.org/10.14686/buefad.732132

Valizadeh, M. (2019). EFL teachers' writing assessment literacy, beliefs, and training needs in the context of Turkey. *Advances in Language and Literary Studies, 10*(6), 53–62. Retrieved from: https://journals.aiac.org.au/index.php/alls/article/view/6040/4302

Vogt, K. & Tsagari, D. (2014) Assessment literacy of foreign language teachers: findings of a European study. *Language Assessment Quarterly, 11*(4), 374-402. https://doi.org/10.1080/15434303.2014.960046

Woodrow, L. (2006). Anxiety and speaking English as a second language. *RELC journal*, *37*(3), 308-328. https://doi.org/10.1177/0033688206071315

Wang, B. (2010). On rater agreement and rater training. English Language Teaching, 3(1), 108-112. OI:10.5539/elt.v3n1p108

Wang H. (2008). A Study on Raters' Interpretation and Application of the Rating Criteria in TEM4-Oral. *Theory and Practice of Foreign Languages Teaching* 2:33-39.

Weigle, S. C. (2002). *Assessing writing.* Cambridge, UK: Cambridge University Press.

Weir, C.J. (1990). *Communicative language testing*. New Jersey: Prentice Hall.

Weir, C.J. (1993). *Understanding and developing language tests*. NewYork: Prentice Hall.

Weir, C. J. (2005). *Language testing and validation*. Hampshire: Palgrave McMillan.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing, 10(3),* 305–319. https://doi.org/10.1177/026553229301000306

Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. Language Testing, 14(1), 85-106. https://doi.org/10.1177/026553229701400105

Wigglesworth, G. and Frost, K. (2017). Task and performance-based assessment. In E. Shohamy and N.H. Hornberger (Eds.), Encyclopedia of language and education:

language testing and assessment (Third Edition, 121-134). Cham, Switzerland: Springer International Publishing.

Wolfe, E.W. (2004). Identifying rater effects using latent trait models. *Psychology Science, 46(1),* 35–51. Retrieved from: https://www.researchgate.net/publication/228986073_Identifying_rater_effects_using_latent_trait_models

Wolfe, E. W. & McVay, A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational Measurement: Issues and Practice, 31*(3), 31-37. https://doi.org/10.1111/j.1745-3992.2012.00241.x

Xobin. (2024). *Psychometric properties of a test: Understanding reliability and validity in assessments*.

Xu, J., Jones, E., Laxton, V., & Galaczi, E. (2021). Assessing L2 English speaking using automated scoring technology: Examining automarker reliability. *Assessment in Education: Principles, Policy & Practice, 28*(4), 411–436. https://doi.org/10.1080/0969594X.2021.1979467

Yıldırım, A. (1999). Nitel araştırma yöntemlerinin temel özellikleri ve eğitim araştırmalarındaki yeri ve önemi. *Eğitim ve Bilim, 23*(112). Retrieved from: https://eb.ted.org.tr/index.php/EB/article/view/5326/1485

Yıldırım, A., & Şimşek, H. (1999). *Sosyal bilimlerde nitel araştırma yöntemleri*. Seçkin Yayınevi.

Yin, R. K. (2014). *Case study research: Design and methods* (5th ed.). Thousand Oaks, CA: Sage.

Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics, 24*(1), 1-27. https://doi.org/10.1093/applin/24.1.1

Zamanzadeh, V., Rassouli, M., Abbaszadeh, A., Nikanfar, A. R., Alavi-Majd, H., & Ghahramanian, A. (2015). *Details of content validity and objectifying it in instrument development*. Nursing Practice Today, 1(3), 163–171. DOI: 10.15171/jcs.2015.017

Zhang, Y., & Elder, C. (2009). Measuring the speaking proficiency of advanced EFL learners in China: The CET–SET solution. *Language Assessment Quarterly, 6*(4), 298–314. https://doi.org/10.1080/15434300902990967

APPENDICES

Appendix 1. Permission from Sakarya University Ethics Committee

Evrak Tarih ve Sayısı: 11.10.2024-E.409292

T.C.
SAKARYA ÜNİVERSİTESİ REKTÖRLÜĞÜ
Etik Kurulu

Sayı : E-61923333-050.99-409292 11.10.2024
Konu : 36/12 Muhammed AK

Sayın Muhammed AK

İlgi : 03.10.2024 tarihli ve E--050.99-0 sayılı yazınız.

Üniversitemiz Eğitim Araştırmaları ve Yayın Etik Kurulu'nun 09.10.2024 tarihli ve 36 sayılı toplantısında alınan "12" nolu karar ile Muhammed AK'ın başvurusu **uygun** görülmüş ve karar örneği ekte sunulmuştur.

Bilgilerinizi rica ederim.

Prof. Dr. Canan LAÇİN ŞİMŞEK
Eğitim Araştırmaları ve Yayın Etik
Kurulu Başkanı V.

Ek: Karar Yazısı (1 Sayfa)

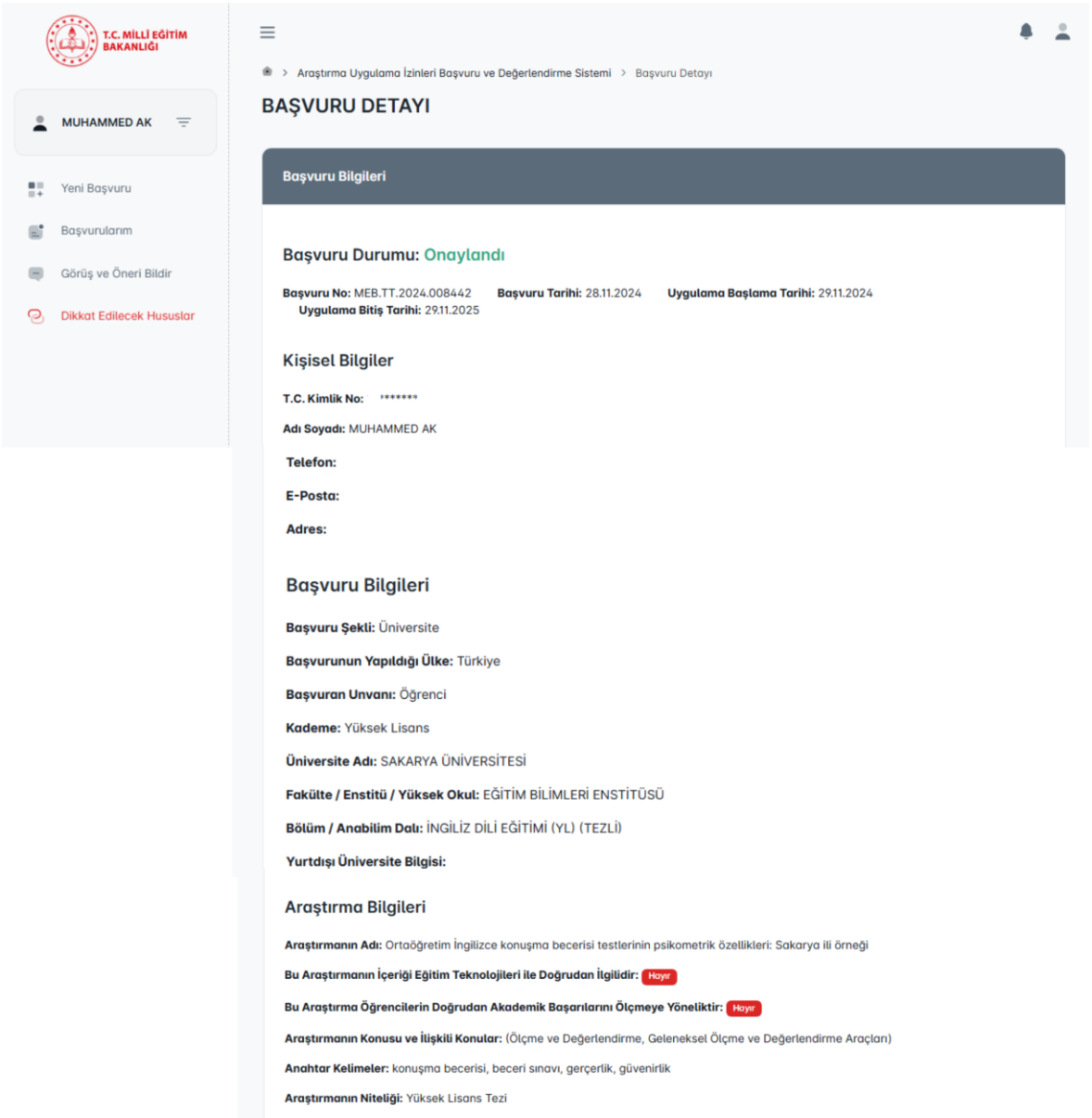
Bu belge, güvenli elektronik imza ile imzalanmıştır.
Doğrulama Kodu :BSDLS9KDK7 Pin Kodu :59862 Belge Takip Adresi : https://turkiye.gov.tr/ebd?eK=5783&eD=BSDLS9KDK7&eS=409292
Adres:Esentepe Kampüsü 54187 Serdivan SAKARYA / KEP Adresi: sakaryauniversitesi@hs01.kep.tr
Telefon No:0264 295 50 00 Faks No:0264 295 50 31
e-Posta:ozelkalem@sakarya.edu.tr Elektronik Ağ:www.sakarya.edu.tr
Bilgi için: Hanife Babacan
Unvanı: Birim Evrak Sorumlusu

KARAR

12. Muhammed AK'ın " Ortaöğretim İngilizce Konuşma Becerisi Testlerinin Psikometrik Özellikleri: Sakarya İli Örneği/ Psychometric Properties of High School English Speaking Tests: The Case of Sakarya " başlıklı çalışması görüşmeye açıldı.

Yapılan görüşmeler sonunda Muhammed AK'ın " Ortaöğretim İngilizce Konuşma Becerisi Testlerinin Psikometrik Özellikleri: Sakarya İli Örneği/ Psychometric Properties of High School English Speaking Tests: The Case of Sakarya " başlıklı çalışmasının Etik açıdan **uygun** olduğuna oy birliği ile karar verildi.

Appendix 2. Permission from the Ministry of National Education



T.C. MİLLİ EĞİTİM BAKANLIĞI

MUHAMMED AK

Yeni Başvuru

Başvurularım

Görüş ve Öneri Bildir

Dikkat Edilecek Hususlar

Araştırma Uygulama İzinleri Başvuru ve Değerlendirme Sistemi > Başvuru Detayı

BAŞVURU DETAYI

Başvuru Bilgileri

Başvuru Durumu: Onaylandı

Başvuru No: MEB.TT.2024.008442 **Başvuru Tarihi:** 28.11.2024 **Uygulama Başlama Tarihi:** 29.11.2024
Uygulama Bitiş Tarihi: 29.11.2025

Kişisel Bilgiler

T.C. Kimlik No: *******

Adı Soyadı: MUHAMMED AK

Telefon:

E-Posta:

Adres:

Başvuru Bilgileri

Başvuru Şekli: Üniversite

Başvurunun Yapıldığı Ülke: Türkiye

Başvuran Unvanı: Öğrenci

Kademe: Yüksek Lisans

Üniversite Adı: SAKARYA ÜNİVERSİTESİ

Fakülte / Enstitü / Yüksek Okul: EĞİTİM BİLİMLERİ ENSTİTÜSÜ

Bölüm / Anabilim Dalı: İNGİLİZ DİLİ EĞİTİMİ (YL) (TEZLİ)

Yurtdışı Üniversite Bilgisi:

Araştırma Bilgileri

Araştırmanın Adı: Ortaöğretim İngilizce konuşma becerisi testlerinin psikometrik özellikleri: Sakarya ili örneği

Bu Araştırmanın İçeriği Eğitim Teknolojileri ile Doğrudan İlgilidir: Hayır

Bu Araştırma Öğrencilerin Doğrudan Akademik Başarılarını Ölçmeye Yöneliktir: Hayır

Araştırmanın Konusu ve İlişkili Konular: (Ölçme ve Değerlendirme, Geleneksel Ölçme ve Değerlendirme Araçları)

Anahtar Kelimeler: konuşma becerisi, beceri sınavı, gerçerlik, güvenirlik

Araştırmanın Niteliği: Yüksek Lisans Tezi

Appendix 3. Informed Consent Form

T.C. Sakarya Üniversitesi

Etik Kurulu

BİLGİLENDİRİLMİŞ GÖNÜLLÜ ONAM FORMU

Sizi Muhammed Ak tarafından yürütülen "Ortaöğretim İngilizce konuşma becerisi testlerinin psikometrik özellikleri: Sakarya ili örneği" başlıklı araştırmaya davet ediyoruz. Bu araştırmanın amacı ortaöğretim düzeyinde konuşma becerisini ölçmek amacıyla uygulanmış sınavların geçerlik ve güvenirlik düzeyini belirlemektir. Araştırmada sizden tahminen on beş dakika ayırmanız ve sınav bilgi formunda istenen bilgileri sağlamanız istenmektedir. Araştırmaya sizin dışınızda tahminen 11 kişi katılacaktır. Bu çalışmaya katılmak tamamen **gönüllülük** esasına dayanmaktadır. Çalışmanın amacına ulaşması için sizden beklenen, bütün soruları eksiksiz, kimsenin baskısı veya telkini altında olmadan, size en uygun gelen cevapları içtenlikle verecek şekilde cevaplamanızdır. Bu formu okuyup onaylamanız, araştırmaya katılmayı kabul ettiğiniz anlamına gelecektir. Ancak, çalışmaya katılmama veya katıldıktan sonra herhangi bir anda çalışmayı bırakma hakkına da sahipsiniz. Bu çalışmadan elde edilecek bilgiler tamamen araştırma amacı ile kullanılacak olup kişisel bilgileriniz **gizli tutulacaktır**; ancak verileriniz yayın amacı ile kullanılabilir. İletişim bilgileriniz ise sadece izninize bağlı olarak ve farklı araştırmacıların sizinle iletişime geçebilmesi için "ortak katılımcı havuzuna" aktarılabilir. Eğer araştırmanın amacı ile ilgili verilen bu bilgiler dışında şimdi veya sonra daha fazla bilgiye ihtiyaç duyarsanız araştırmacıya şimdi sorabilir veya e-posta adresi ve numaralı telefondan ulaşabilirsiniz. Araştırma tamamlandığında genel/size özel sonuçların sizinle paylaşılmasını istiyorsanız lütfen araştırmacıya iletiniz.

Yukarıda yer alan ve araştırmadan önce katılımcıya verilmesi gereken bilgileri okudum ve katılmam istenen çalışmanın kapsamını ve amacını, gönüllü olarak üzerime düşen sorumlulukları anladım. Çalışma hakkında yazılı ve sözlü açıklama aşağıda adı belirtilen araştırmacı/araştırmacılar tarafından yapıldı. Bana, çalışmanın muhtemel riskleri ve faydaları sözlü olarak da anlatıldı. Kişisel bilgilerimin özenle korunacağı konusunda yeterli güven verildi.

Bu koşullarda söz konusu araştırmaya kendi isteğimle, hiçbir baskı ve telkin olmaksızın katılmayı kabul ediyorum.

Katılımcının
Adı-Soyadı:.....
İmzası: İletişim Bilgileri: e-posta: Telefon:
İletişim bilgilerimin diğer araştırmacıların benimle iletişime geçebilmesi için "ortak araştırma havuzuna" aktarılmasını; ☐ Kabul ediyorum ☐ Kabul etmiyorum (lütfen uygun seçeneği işaretleyiniz)

Velayet veya Vesayet Altında Bulunanlar İçin;

Veli veya Vasisinin

Adı-Soyadı:.....
İmzası:
Araştırmacının

Adı-Soyadı: Muhammed Ak
İmzası:

Şahidin:[1]
Adı-Soyadı:.....
İmzası:

[1]Şahit Kriterleri: Çalışmanın bir üyesi olmayan, araştırmacı tarafından belirlenen ve araştırmanın bulguları üzerinde herhangi bir olumlu/olumsuz etki yaratma olasılığı bulunmayan tarafsız yetişkinlerdir. Katılımcı araştırmaya katılmayı kabul edip onam formunu imzalamayı istemediği durumlarda araştırmacı onam formundaki bilgileri katılımcıya sözlü olarak okur. Katılımcı onayladığını sözlü olarak beyan ederse şahit de bu sözlü onam sürecine yazılı onam formunu imzalamak sureti ile şahitlik ettiğini beyan etmiş olur.

Appendix 4. Exam Information Form

Sınav Bilgileri (10.sınıf I.Dönem II.Sınav) Bildirim Formu

Okul Adı: Sınav Kapsamı(Üniteler):

1.Ünite☐ 2.Ünite☐ 3.Ünite☐ 4.Ünite☐ 5.Ünite☐ Diğer:

Konuşma becerisi ölçümünde kullandığınız yöntemi anlatınız

Ölçme Yöntemi(diyalog, sunum, vb.):

Sınav öncesi yaptıklarınız(duyuru, sınav içeriği belirleme, vb.): Öğrencilere sınav günü sınav öncesi hazırlık için verilen süre(varsa):

Öğrencilerin konuşma becerisini ölçmek için performans göstermeleri için tanınan süre:

Sınav anı için alınan önlemler(ortam, uygulanış, vb.)(varsa):

Sınav Yönergeniz:

Ölçme ve değerlendirme hazırlık sürecinden puanlama sürecinin sonuna kadar aşağıdaki yöntemleri kullanıp kullanmadığınızı belirtiniz.

	Evet	Hayır
Test tekrar test yöntemi uygulaması (Testin aynı bireylere aynı koşullarda farklı zamanlarda uygulanması ve aradaki farka bakılması)		
Paralel (eşdeğer) test yöntemi uygulaması (Benzer özellikte bir testin yapılacak test ile beraber yapılıp ikisi arasındaki farka bakılması)		
Eşdeğer yarılar yöntemi uygulaması (Testi içerik ve zorluk olarak iki yarıya bölüp ikisi arasındaki farka bakılması)		
Çoklu Puanlayıcı (Aynı sınav kağıdını birden fazla kişinin puanlaması ve puanlamalar arasındaki farka bakılması		
Tekrarlı puanlama uygulaması (Aynı sınav kağıdını aynı kişinin farklı zamanlarda en az iki kez puanlaması ve puanlamalar arasındaki farka bakılması		
Dış puanlayıcı uygulaması (Puanlamanın derse giren öğretmen haricinde bir öğretmen tarafından yapılması)		
Puanlamada Rubrik (değerlendirme kriterleri) kullanımı*		

*Puanlama aşamasında rubrik(değerlendirme kriterleri) kullanıldı ise kriterleri ve varsa puan değerlerini belirtiniz veya ek olarak araştırmacı ile paylaşınız.

Konuşma becerisini ölçmek için kullandığınız sınav örneğini araştırmacı ile paylaşınız.

Appendix 5. Expert Evaluation Checklist

İNGİLİZCE KONUŞMA BECERİSİ TESTLERİNİN PSİKOMETRİK ÖZELLİKLERİ UZMAN KONTROL LİSTESİ

Değerli uzman, sizden aşağıdaki tabloyu sınav bilgi formu ve eklerine göre doldurmanız beklenmektedir. Herhangi bir madde ile ilgili yetersiz veri durumunda fikrim yok kısmını işaretleyebilir; eklemek istediklerinizi açıklamalar kısmına ekleyebilirsiniz.

		EVET	KISMEN	HAYIR	FİKRİM YOK	AÇIKLAMA
Görünüş Geçerliği	Ölçme uygulaması ölçülmesi planlanan kazanım ile açık bir şekilde ilişkili görünmektedir.					
	Ölçme uygulaması katılımcıların(lise 9 ve 10. sınıf / 14-16 yaş) yaşı için uygun görünmektedir.					
Kapsam Geçerliği	İlgili sınav kapsamındaki kazanımlar sınava dahil edilmiştir.					
	Ölçme aracı kazanımların bilişsel basamak seviyelerine (bilgi, kavrama, uygulama, analiz, sentez, değerlendirme) uygundur.					
Yapı Geçerliği	Ölçme aracı ölçülmek istenen kazanımları ölçecek şekilde tasarlanmıştır.					
	Ölçme aracı ölçülmek istenen kazanımların dışında bir bilgi-beceri gerektirmemektedir.					
	Ölçme uygulaması olası yanlılığa karşı önlem içermektedir.					
	Ölçme uygulamasında amaca uygun açık ve detaylı bir puanlama rubriği kullanılmıştır.					
Güvenirlik	Sınavın açık, net ve anlaşılır bir yönergesi vardır.					
	Sınavın içeriği açık net ve anlaşılır düzeydedir.					
	Puanlamada kullanılan kriterler amaca ve içeriğe uygundur.					
	Puanlayıcı güvenirliği sağlanmıştır.					
	Sınavın belirtilen uygulanış şekli amaca ve içeriğe uygundur.					