Washback Effect of a National English Language Teacher Field Knowledge Test for Teacher Recruitment in Türkiye: A Scale Development Study

Demet ÜNAL AYDIN[1] & Sevgi ŞAHİN[2]

[1]Atılım University, Ankara, TÜRKİYE
demet.unal@atilim.edu.tr
https://orcid.org/0009-0009-1715-0148

[2]Ph.D., Başkent University, Ankara, TÜRKİYE
ssahin@baskent.edu.tr
https://orcid.org/0000-0002-8385-2756

Abstract

This study aims to offer a validated scale to examine the washback effect of a high-stakes test, a national English Language Teacher Field Knowledge Test (TFKT) for EFL teacher recruitment in Türkiye, on pre-service EFL teachers' professional development and EFL teacher education programs, along with their awareness and perception of TFKT. To this end, after a thorough literature review, researchers created and validated a new scale. Data were obtained from pre-service EFL teachers preparing for TFKT in two phases. First, data were collected from 195 participants for the pilot study, and the scale's reliability and validity were tested by Cronbach's Alpha Analysis and Exploratory Factor Analysis. After the necessary revisions, the scale was administered to another group of informants (n.300), and the reliability was checked a second time. The findings emphasized that the scale was indeed reliable and valid, preparing the groundwork for further research. As a result, this research provides a reliable assessment tool for evaluating the impact of TFKT on pre-service EFL teachers and EFL teacher education. The established scale shows potential for guiding educational policies and practices with substantial contributions to the ongoing enhancement of EFL teacher education programs.

Keywords: EFL Teacher education; Teacher field knowledge test; Washback effect; High-stakes tests; Scale development; Language assessment; Pre-service EFL teachers

İngilizce Öğretmenlik Alan Bilgisi Testi'nin Etkisi: Ölçek Geliştirme Çalışması

Özet

Bu çalışma, Yüksek Öğretim Kurulu tarafından Türkiye'de İngilizce öğretmeni alımı için geliştirilen ve ulusal bir sınav olan İngilizce Öğretmenlik Alan Bilgisi Testi' nin (ÖABT) Türkiye'deki hizmet öncesi İngilizce öğretmenleri ve İngiliz dili eğitimi programları üzerindeki etkisini ve bununla birlikte sınavla ilgili farkındalık ve algılarını incelemeyi amaçlamaktadır. Bu amaçla, kapsamlı bir literatür taraması ile, araştırmacılar tarafından yeni bir ölçek oluşturulmuş ve geçerlik ve güvenilirliği saptanmıştır. Veriler, iki aşamada ÖABT'ye hazırlanan İngilizce öğretmen adaylarından toplanmıştır. İlk olarak pilot çalışma için 195 katılımcıdan veri toplanmış, ölçeğin güvenilirliği ve geçerliliği Cronbach's Alpha Analizi ve Açımlayıcı Faktör Analizi ile test edilmiştir. Gerekli revizyonlardan sonra, ölçek 300 katılımcıya daha uygulanmış ve güvenirliği ikinci kez kontrol edilmiştir. Bulgular, ölçeğin güvenilir ve geçerli olduğunu vurgulayarak daha fazla araştırmaya zemin hazırlamıştır. Bu araştırma, ÖABT'nin hizmet öncesi İngilizce öğretmenleri ve İngilizce öğretmen eğitimi üzerindeki etkisini değerlendirmek ve ÖABT ile ilgili farkındalık ve algılarını araştırmak için güvenilir ve geçerli bir değerlendirme aracı sunmaktadır. Oluşturulan ölçek, İngiliz Dili Eğitimi programlarının geliştirilmesi ve eğitim politikalarına ve uygulamalarına rehberlik etme potansiyeline sahiptir.

Anahtar Sözcükler: İngilizce Öğretmenlik alan bilgisi testi (ÖABT); Sınavın ket vurma etkisi; Ölçek geliştirme; İngilizce öğretmen adayları; Öğretmen Yetiştirme

1. Introduction

Every country is responsible for providing exemplary instruction, promoting successful learning experiences, and enforcing effective evaluation procedures within English as a Foreign (EFL) Language teacher education. These components are crucial in determining the recruitment process for English language teachers, guaranteeing they possess the necessary teaching skills and knowledge to manage the diverse linguistic and educational environments they will face in their classrooms. Globally, the quality of EFL teachers is significantly influenced by teacher education and recruiting processes. Therefore, undoubtedly, the relationship between a country's education system and its evaluation of teachers for recruitment is interwoven (Johnstone, 2004), which means teacher recruitment has an essential impact on the overall education system of a nation. The procedure of selecting and recruiting teachers through careful and appropriate assessment of their knowledge base is critical to ensure an adequate supply of competent and efficient instructors (Allen, 2005).

EFL teachers' competency and efficiency are strongly linked to the Teacher Knowledge Base (TKB). Shulman (1987) divides the knowledge base into content knowledge, general pedagogical knowledge, curriculum knowledge, pedagogical content knowledge, learner characteristics, educational contexts, and educational goals and values. From the standpoint of foreign language teachers, the expected knowledge base differs. The value of subject matter expertise (Thornbury, 1997), or subject matter content knowledge (Shulman, 1987), is a recurring focal point in discussions on teacher language awareness. This aspect of knowledge relates to a good command of the English language in terms of phonology, semantics, syntax, pragmatics, as well as literary and cultural perspectives. EFL teachers are required to skillfully manage and combine these elements to promote inclusive language learning experiences. As another component of TKB, Pedagogical Knowledge (PK) refers to teachers' familiarity with teaching tactics and procedures, as well as their familiarity with students' cognitive, social, and emotional growth (Shulman, 1987). The last component of TKB, Pedagogical Content Knowledge (PCK), lays the foundation for effective and appropriate teaching and assessment of the content knowledge. For a thorough evaluation and recruitment of foreign language teachers, assessments of teachers for recruitment should reflect a balance of content, pedagogical content, and pedagogical knowledge.

At this point, it is safe to state that any high-stakes tests, such as those for teacher recruitment, would be associated with washback research due to their critical roles and possible consequences in people's educational and professional lives. This complex, multi-faceted concept (Alderson, 2004), which has been subject to many interpretations throughout history, acts as a central point for examining the broader implications of educational assessments. Washback, as defined by Messick (1996), means the impact of a test on language development and the resulting influence on teachers' adoption of non-standard techniques, which can either enhance or impede language learning. That is, washback may include purposeful or unforeseen curricular change resulting from a shift in public evaluations because certain test characteristics might induce washback (Cheng, 2005). Numerous scholars have concluded that washback appears to be predominantly linked to high-stakes examinations, which primarily determine significant societal matters, including education and the economy (Pearson, 1988; Shohamy, 1993). Therefore, compared to lower-stakes examinations, high-stakes tests have more significant impacts with a higher potential to induce positive or negative and intended or unintended consequences (Sadeghi et al., 2021) and changes in instructional practices and strategies (Shohamy, 2017) because the outcomes are utilized to determine critical stakeholder decisions (Thomas et al., 1998).

One example of such a high-stakes test is the English Language Teacher Field Knowledge Test (TFKT) administered for EFL teachers' evaluation and recruitment purposes to appoint teachers to public primary, secondary, and high schools in Türkiye. TFKT is a standardized test candidate that EFL teachers should take in addition to two separate paper-based tests known as the Personnel Placement and Selection Exam (PPSE, i.e., KPSS) in order to be recruited for public schools by the Ministry of Nation Education (MoNE). Candidates for English teaching positions who take the PPSE are assessed in three separate paper-based tests in three different sessions: general aptitude and world knowledge, knowledge of educational sciences, and field knowledge test (TFKT). Evidently, because of its nationwide application and critical role in teacher recruitment, it is a high-stakes test that is naturally expected to have varying types and degrees of washback effects. Despite this, it has not caught the attention of many scholars, to the best of the researchers' knowledge, in terms of evaluating its washback effect, although there are some studies conducted to investigate pre-service EFL teachers' opinions and perceptions of TFKT

(Elmacı, 2015; Karaer et al., 2018; Karataş & Okan, 2021; Sert, 2015), pre-service EFL teachers' opinions regarding strengths and weaknesses of TFKT (Sert, 2018), and pre-service EFL teachers opinions related to the current teacher recruitment process and model in Türkiye in general (Erdoğan, 2019; Yeşilçınar & Çakır, 2020). In Türkiye, a highly exam-oriented county (Hatipoğlu, 2010) where such a high-stake test is administered to thousands of candidates each year to determine if they are eligible to teach in public schools, extensive research must be conducted to reveal its consequences so that it could lead to refinements in the recruitment model if necessary. Thus, the researchers aimed to set the groundwork for further washback research on TFKT by designing and establishing a reliable and valid scale to investigate the washback impact of the test. Moreover, it was also the intention of the researchers that an in-depth investigation into the pre-service EFL teachers' awareness of and perception on TFKT is enabled through this scale. To this end, this study aims to answer the following questions:

1. What are the results of the scale's reliability?

2. What are the results of the scale's validity?

2. Literature Review

Washback, also known as backwash, refers to the influence that a test has on the process of teaching and learning (Alderson & Wall, 1993; Biggs, 1995). Positive (beneficial) or negative (harmful) are the two predominant perceptions regarding washback. Negative washback occurs when the content or format of a test is predicated on a limited conception of language proficiency, thereby imposing limitations on the educational environment (Taylor, 2005). Positive washback, however, occurs when a testing procedure incentivizes "good" teaching practice; for instance, implementing an oral proficiency test to enhance speaking skill instruction is considered positive washback (Taylor, 2005). In this context, it is argued by Messick (1996) and Hughes (2004) that there is a clear and strong connection between the teaching process and the design and utilization of the test since a poorly planned exam can potentially have negative results, whereas a well-prepared test can positively influence the teaching and learning process (Frederiksen,1984; Hughes, 2004).

In recent years, there has been a notable surge in research focused on language testing and assessment to examine the complex characteristics and underlying mechanisms of washback effects of language testing on various facets of EFL education and instruction. Extensive evidence supports the notion that assessments, particularly those that are high-stakes, induce substantial washback effects on instruction and learning across various academic settings (Andrews et al., 1997; Burrows, 1999; Cheng, 1999; Scaramucci, 2002; Shohamy et al., 1996; Watanabe, 1996). Nevertheless, the magnitude of these effects varies among individuals and distinct facets of instruction and learning within a particular academic setting (Alderson & Hamp-Lyons, 1996).

That being said, EFL teacher education programs are one of such academic settings mentioned. In all countries, certain types of evaluation are in place for teacher candidates to be admitted to EFL teacher education programs. After admission, they are educated on the multifaceted nature of teaching English throughout their pre-service training. Finally, when they graduate, they are evaluated once again to be recruited as English teachers. In some countries (e.g., Finland), this evaluation is carried out by committees or boards interviewing the candidates, whereas in others, such as Türkiye, standardized high-stakes assessments are utilized. As stated by Cheng (2005), the outcomes of high-stakes tests are used to perform crucial washback impacts on test takers' future academic and job opportunities. They can have significant effects on the test takers. Therefore, this is a highly significant area to be examined.

Scrutiny of the literature shows that there is no universally applied method of assessing and recruiting qualified individuals to teach a foreign language at educational institutions. For instance, a decentralized approach is on display in the United States regarding the methods used to recruit EFL teachers, which differ among states and districts (Goldhaber & Hansén, 2010; Nettles et al., 2011). When hiring EFL instructors in the United States, the Praxis exams, specifically Praxis II, come into play in most states. Praxis II, also known as Praxis Subject Assessments, guarantee that educators are adequately equipped to impart high-caliber instruction to students by evaluating knowledge and pedagogical abilities in K–12 subjects. Praxis Subject Assessment for EFL teachers includes 130 selected-response questions on content knowledge. There has been some research on Praxis II in the foreign language teaching context to see whether it meets program criteria and how it helps language teachers grow professionally,

as a result of which this test has changed a lot and is based on criteria set by educational groups. Therefore, it impacts how language instructors in the US are prepared and hired (Moser, 2012). However, no research has been conducted on the washback effect of Praxis tests in the EFL context.

In Thailand, however, the nation's recruitment procedure incorporates an assessment after pre-service training for teacher candidates to complete successfully. Teaching and Educational Aptitude Test (TPAT 5), a nationwide standardized test, is implemented to recruit new teachers, and the scores obtained from this test are significant factors in the selection procedure. This test is composed of a multiple choice one based on content knowledge prepared and administered by the National Institute of Educational Testing Service. Considering the test scores, the Ministry of National Education (MoNE) oversees a unified system through which educators are hired and compensated. Still, the washback impact of TPAT 5 has not been examined so far.

India is another example of a country using a standardized examination to recruit teachers. Just like in the USA, India has a state system. Therefore, each state oversees the recruitment process by itself. In Sikkim, for instance, the State Teacher Eligibility Test (STET) is employed to ensure a fair, inclusive, and meaningful selection of teacher candidates. There are two test stages, which are separated according to which grades will be taught after recruitment. Candidate teachers who wish to teach grades 1 to 5 take Paper I as a test, whereas those who intend to teach grades 6 to 8 take Paper II. Both papers include multiple-choice items covering a range of areas, such as child development and pedagogy, language proficiency, language acquisition, and language teaching, with a total of 150 questions. As a high-stakes test utilized for the recruitment of teachers, STET also remains unstudied in terms of its washback effect.

Similarly, in Türkiye, pre-service EFL teachers are evaluated not only before admission to EFL teacher education programs with a nationwide high-stakes test but also after graduation with another high-stakes test (TFKT) to be recruited in public schools. However, once again, the literature review related to the studies on EFL teacher evaluation and recruitment in Türkiye reveals that there is a scarcity of research on the Public Personnel Selection Examination (PPSE, i.e., KPSS) and TFKT. The existing research, however, focuses on pre-service EFL teachers' perceptions and opinions regarding PPSE-TFKT in general.

In the realm of high-stakes tests, such as public examinations, the washback effect extends to influencing the attitudes, behaviors, and motivation of instructors, learners, and their families (Pearson, 1988). Its perception as positive or negative depends on various factors, including who conducts the research, the educational setting, the timing, the duration and frequency of assessment methods, the purpose, and how people utilize such assessment instruments in their settings (Cheng & Curtis, 2004). While washback has been researched in the field of EFL education, including numerous studies in Türkiye (Akpınar & Çakıldere, 2013; Çakır, 2017; Hatipoğlu, 2016; Kılıçkaya, 2016; Külekçi, 2016; Özmen, 2011a; Özmen, 2011b; Sayın & Arslan, 2016; Yeşilçınar, 2018; Yıldırım, 2010), studies focusing on the washback effect of TFKT in particular seem to be absent. In the studies conducted on washback, the results pointed to negative washback, such as in Hatipoğlu's (2016) study on English Section of the University Entrance Exam (ESUEE) concluding that the existence of the test alone has a significant impact on the way English is taught and learned in Türkiye, including its planning, definition, and structure.

For instance, Karaer et al. (2018) examined prospective teachers' views on TFKT conducting a survey with 306 teacher candidates from 16 disciplines. The results revealed that teacher applicants felt that their graduation marks were sufficient to be appointed to public schools and chose not to go through any further assessment procedure. With a similar approach, Sert (2015) investigated test takers' opinions to analyze the TFKT. The researcher collected data from 34 pre-service EFL teachers via a survey that included open-ended questions. The results showed that participants held positive attitudes toward the test since the test allowed for their appointment. Karataş and Okan (2021) conducted a case study to examine the roots, consequences, and ramifications of PPSE-TFKT. The results revealed that the PPSE-TFKT resulted in several unexpected effects and implications, such as the society attributing a decisive role to the test and judging the test-takers' value based on this role. However, it is essential to note that the study's questions were restricted to PPSE rather than TFKT. Although the findings point to washback, the primary intention was to investigate the power of the test rather than its washback impact.

In another study, Yağcı and Kurşunlu (2017) examined how potential teachers conceive the connection between PPSE-TFKT and the course design in the faculties of education. The findings showed that

prospective teachers believed their programs provided inadequate professional education. In preparation for the exam, aspiring instructors reported 'moderate' competency and a mismatch between test skills and teacher education program content. They believed the test was faulty and subjective, and it was also stated that it was not a valuable tool for identifying qualified instructors. In line with these findings, when Atav and Sönmez (2013) examined PPSE-TFKT's importance for pre-service teachers, they reached a somewhat similar conclusion in their research. The analysis of the data collected from 300 participants from various education programs uncovered that test-takers thought their program curriculum was inconsistent with the exam subject. Thus, they had to take private courses to pass. The survey also found that pre-service teachers believed the test had an adverse impact on their lives and undergraduate education. They also considered the exam unsuitable for hiring skilled instructors as a multiple-choice test. They advised adding an oral and applied portion and repeating it after the teacher appointment by MONE. As can be seen, although some impacts of the test were listed among the findings, none of the studies mentioned explicitly focused on the washback effect of TFKT.

Despite the expanding body of literature concerning washback, empirical research in this domain remains relatively scarce, especially in language teacher education and evaluation. The extent to which context and washback are interdependent and the circumstances under which different forms of washback are most likely to be induced by testing in a given educational setting remain challenging to predict. To advance our comprehension of this phenomenon, it is imperative to examine it within a particular educational context through comprehensive investigations of various facets of teaching and learning. Surprisingly, despite the significance of these concepts in EFL teacher education and teacher assessment, a literature gap in washback research has been identified both in Türkiye and abroad. This realization, coupled with the absence of a comprehensive washback scale to evaluate the impact of TFKT on EFL teacher candidates in Türkiye, prompted the research initiative. The researchers aim to provide a valid scale to compare and contrast the relation among these intricate constructs by examining the TFKT's washback effect and EFL teacher candidates' awareness and perception levels regarding the test. Thus, it is believed the study will fill this niche in the literature and contribute valuable insights into the impact of this high-stakes test, TFKT, on EFL teacher candidates and EFL teacher education programs, laying the groundwork for measuring washback not only in the Turkish context but also abroad.

3. Method

As this is a scale development study, the researchers employed a quantitative research design, a methodical and objective technique to generate knowledge. It relies on numerical data, statistical analysis, and rigorous study methodologies to examine correlations between variables and evaluate hypotheses (Creswell, 2009). In the scale development procedure, adopting phases from various sources, the following steps were taken: determining the purpose of the scale (DeVellis & Thorpe, 2021), determining to whom and why it will be applied (Seçer, 2018), deciding on the extent and content of the scale (DeVellis & Thorpe, 2021); generating items following the previously set extent and content (Şeker & Gençdoğan, 2006); item control and creating a scale form (DeVellis & Thorpe, 2021); defining the scoring procedure of the items and how to analyze the data (Cohen et al., 2013); applying the scale to be developed in the scale development group (DeVellis & Thorpe, 2021); scoring and analyzing items; and finalizing the scale in line with the outcomes achieved (Crocker & Algina, 1986) (Please See Ünal Aydın, 2024 for an in-depth presentation of the research design).

3.1. Research Context

The education system in Türkiye places a strong emphasis on examinations. The evaluation of the success of students, instructors, and schools is based on the student's proficiency in several examinations (Hatipoğlu, 2016), and TFKT is a standardized national high-stakes examination administered by the Student Selection and Placement Center (i.e., ÖSYM) for EFL teacher candidates. Until 2013, the appointment of teacher candidates according to PPSE was carried out depending on the results of the Educational Sciences Test (EST), General Culture Test (GCT), and General Aptitude Test (GAT). So, EFL teacher candidates who took the PPSE were assessed in two separate paper-based tests in two different sessions: general aptitude and world knowledge and knowledge of educational sciences. The first test comprises multiple-choice questions about current events, geography, Turkish, mathematics, and history. In contrast, the second one is an afternoon session of educational sciences given on the same day. Then, in October 2013, in addition to these tests, the Teacher Field Knowledge Test (TFKT, i.e., ÖABT)

was introduced, probably due to previous research stressing the candidates' need for a field test. This test, scheduled for a separate day two weeks after the first two tests are administered, aims to evaluate EFL teacher candidates' content knowledge, pedagogical knowledge, and pedagogical content knowledge. Thus, with the arrival of the new examination, TFKT contributes 50%, EST contributes 20%, GCT and GAT contribute 30% to their overall scores (MoNE, 2015). In 2016, an interview was also brought forth in addition to these tests for teacher appointments. In this system, in the case of EFL teachers' appointments, eligibility to take the exam is related to graduating from faculties of education or faculties of science and letters with the prerequisite of holding a Pedagogical Formation Certificate (MoNE, 2015).

As the final step of the teacher evaluation process, EFL teachers who get passing scores on the test are invited to an oral interview in which a panel of three people evaluates candidates on a variety of factors, including their ability to understand and summarize complex ideas, express themselves clearly, communicate effectively, be open to new ideas and willing to incorporate technological advances, and so on (MoNE, 2015). Candidates who achieve a minimum score of 60 in the oral interview phase are given the opportunity to be employed at their preferred schools based on their rankings (MoNe, 2015)

TFKT aims to ensure that only the most skilled and capable teachers are selected for teaching positions in the Turkish education system. It might have originated in Türkiye as a response to the need for a consistent and unbiased approach to assessing pre-service teachers' content knowledge and teaching skills. Given the increasing significance of foreign language education in the country and the escalating need for well-educated teachers, it became imperative to create a standardized tool for assessing the classroom readiness of teacher candidates. So, it can be stated that the idea of creating TFKT was driven by the imperative to guarantee that teachers had the necessary knowledge base to effectively educate when they come into duty. Since TFKT coverage is based on the curriculum of English Language Teacher Education (ELTE) programs (i.e., the 4-year undergraduate program at the Department of Foreign Languages in Education Faculties in Türkiye), the test assesses EFL teacher candidates' language proficiency, content knowledge, and pedagogical content knowledge. Table 1 below displays the detailed coverage of the test.

Table 1.

The Detailed Content of TFKT

English Language Teaching & TFKT Content	**Overall Percentage**	**Approximate Number of Items in TFKT**	**Time Allocated**
Field Knowledge Test	60%	45	
1 Language Proficiency (Cloze Test &Paragraph Questions)		25	
2 Linguistics		10	
3 Literature		10	
Field Educational Knowledge Test	40%	30	
1 Approaches, Methods, and Techniques in ELT		4	
2 Teaching Language Skills in ELT		14	120 minutes
3 Teaching English to Young Learners		3	
4 Materials Development, Adaptation, and Evaluation		4	
5 Language Testing and Assessment		2	
6 Language Acquisition		3	
Total	100%	75	

3.2. Participants

For this study, data were collected from two groups of EFL teacher candidates. The first group of participants consisted of 195 EFL teacher candidates (173 female and 22 male respondents) from 19 different universities. The age range was from 20 to 42. Furthermore, as TFKT is available to individuals

who have completed language-related programs, participants were either enrolled in (n.21) or had completed such programs (n. 174), provided they met the Pedagogical Formation prerequisite. The more detailed descriptive statistics about the programs the participants majored in can be seen in Table 2.

Table 2.

Demographic Information of the First Sample

Gender	**Frequency**	**Percent**
Male	22	11,28
Female	173	88,72
Program		
English Language Teaching	101	51,8
English Language and Literature	80	41,0
American Culture and Literature	3	1,5
English Linguistics	6	3,1
Translation Studies	5	2,6
Educational Status		
Senior (4th year) student	21	10,8
Graduate	174	89,2
TFKT Taking Status		
Yes	168	86,15
No	27	13,85
Preference for Appointment to Public Schools		
Yes	165	84,62
No	11	5,64
Not Sure	19	9,74

The participants were also queried regarding their prior TFKT experience, as the presence or absence of such knowledge could potentially influence their understanding and perception of the examination. It was found that 168 respondents took the examination before, while 27 did not. The participants were also asked whether they would like to be appointed as EFL teachers to public schools in Türkiye as a component of the demographic information. The findings revealed that an overwhelming majority of the respondents (n=165) expressed a desire to be appointed.

For the second round of data collection, 300 EFL teacher candidates from 24 different universities in Türkiye (235 females and 65 males) filled out the scale. Their ages ranged from 20 to 54. Of 300 participants, 201 were graduates (173 from ELT Programs, 112 from ELL, and 15 from other programs), and the remaining 99 were still in their senior years of university. Moreover, 180 respondents had already taken the test at least once, whereas 120 would take the test for the first time. Finally, when asked if they would like to be appointed to public schools, 23 responded "No," 43 were undecided about the appointment, while the rest (n=234) expressed their desire to be appointed. Overall, it can be stated that 78% of the participants were motivated to be appointed to public schools by receiving a sufficient score from TFKT. Table 3 presents the necessary demographic information for the second sample group.

Table 3.

Demographic Information of the Second Sample

Gender	**Frequency**	**Percent**
Male	65	21,7
Female	235	78,3
Program		
English Language Teaching (ELT)	173	57,7
English Language and Literature (ELL)	112	37,3
American Culture and Literature	5	1,7
English Linguistics	7	2,3
Translation Studies	3	1,0
Educational Status		

Senior (4th year) student	99	33,0
Graduate	201	67,0
TFKT Taking Status		
Yes	180	60,0
No	120	40,0
Preference for Appointment to Public Schools		
Yes	234	78,0
No	23	7,7
Not Sure	43	14,3

3.3. Data Collection

3.3.1. Data Collection Tool Construction

The researchers designed a scale to investigate the washback effect of TFKT on pre-service EFL teachers and EFL teacher education as well as their perceptions on and awareness of the test. After a thorough literature review on washback, EFL teacher education, and evaluation and recruitment in Türkiye and wordwide, the scale's items were written, and a preliminary draft of the scale was completed. Following the construction of the first draft of the data collection tool, five academicians were interviewed to obtain expert opinions. They were specialized in EFL Teacher Education and Language Assessment with varying degrees of experience and expertise in TFKT. To ascertain the alignment of each item in the scale with the research purpose, the experts systematically evaluated the scale's content, item wording, and questionnaire design. The questionnaire was modified following the suggestions and changes offered by the experts to enhance its comprehensibility. With this, content validity was also aimed to ensure so that the items represented the construct measured by the scale (DeVellis & Thorpe, 2021). Finally, the face validity and the content validity of the scale were approved by the same experts after making changes to the scale.

The execution of pilot testing for data collection tool is of the highest significance since it offers a multitude of benefits that improve the overall quality and validity of the study (Audet et al., 2023). Thus, to avoid any problems with the data collection instrument and to check and ensure the construct and concurrent validity, researchers conducted a pilot study to identify and rectify any ambiguous or obscure sections of the scale for the participants.

Data from both groups of participants (i.e., respondents for the pilot study and the main study) were gathered by sharing the link to the survey with universities and social media platforms to reach as many respondents as possible from diverse regions of the country to ensure validity. The data collection process in the pilot study lasted for two weeks, starting in August 2023 and finishing in September 2023. The data for the main study lasted for three months, from September 2023 to November 2023. Overall, the data collection process took a total of 4 months.

3.3.2. Data Collection Tool

The last version of the survey comprised three main parts: demographic information, a 4-point Likert scale, and a single open-ended item.

The introductory part of the survey was specifically crafted to collect demographic data from the participants, which comprised general information such as gender, age, program or graduation status, and so forth. Furthermore, the survey aimed to gather information on the participants' previous encounters with TFKT, if applicable, along with their inclinations concerning admission to public schools (specifically MoNE institutions) in Türkiye.

Furthermore, the Likert scale consisted of three sections:

(1) The first section evaluated the respondents' level of awareness of TFKT through 12 items. The participants were presented with the following options to indicate their level of familiarity: "Definitely Not," "Definitely," "Probably," or "Not Sure."

(2) The second section aimed to determine the participants' perspectives on TFKT via 13 items. In this part, respondents were provided with the following options to express their degree of concurrence: "Strongly Agree," "Agree," "Disagree," or "Strongly Disagree." This was done to evaluate the perceived utility and worth of TFKT.

(3) The third section of the scale assessed the washback effect on the pre-service EFL teachers, EFL teacher education, and the future professional career implications of TFKT employing 21 items.

The concluding part of the survey consisted of a single open-ended inquiry that requested respondents to share their perspectives or recollections regarding previous interactions with TFKT. With this, participants were encouraged to provide any further insights or opinions regarding elements they believe ought to have been considered in the survey.

3.4. Data Analysis

As the primary objective of this study was to determine the validity and reliability of the scale, statistical analyses were conducted using SPSS 27.0.1. In order to ensure the reliability of the scale, Cronbach's alpha test was administered.

In addition, Exploratory Factor Analysis (EFA) was conducted to ascertain the underlying factor structure of the scale, and the researchers, in doing so, aimed to determine if the items on the scale assessed a single underlying construct to prove that the scale had construct validity as well. Factor Analysis is a statistical technique used to analyze a group of variables and identify subsets of variables that are independent from each other (Tabachnick & Fidell, 2013). It is a valuable tool for identifying the underlying factors of variables by grouping similar variables together in the same factor (Verma & Abdel-Salam, 2019). Exploratory factor analysis, on the other hand, is employed to assess the dimensionality of a dataset and is commonly utilized in the first phases of research to get insights into the interconnections between a group of variables (Pituch & Stevens, 2016). Put simply, it assesses the appropriateness of the sample size. For both the overall model and each individual variable, the test determines if the sample is large enough. This is significant since a large sample size can enhance the representativeness of the sample, result in greater statistical power, help generalize findings from the sample to the population and lead to a greater precision in estimating the relationships between the variables.

Furthermore, after the pilot process and the necessary revisions, the scale was applied to another sample of 300 respondents again to ensure reliability. Based on the data collected from this second group, Cronbach's alpha test was administered again.

4. Findings

Cronbach's Alpha Analysis was conducted on the dataset gathered in the pilot test to assess the internal reliability of the questionnaire. The analysis was conducted independently for the initial 12 items and the subsequent 34 items of the scale due to the distinct determiners used for the awareness scale and the perception and washback scale. The Cronbach's Alpha value for the former was determined as .804, indicating a high level of internal consistency. Nevertheless, the analysis also indicated that removing items 1, 6, and 7 from the scale would enhance its reliability (see Table 4). Upon closer analysis of the findings, it was determined that item 1 could be excluded as it was a redundant item whose response could also be derived from items 2, 3, 4, and 5. Furthermore, item 6 was deemed dispensable since it did not directly contribute to addressing any of the study inquiries. Ultimately, it was determined that item 7 should be deleted due to its classification as a double-barrelled item. Upon evaluating the analytical findings, the items in concern were removed, and a subsequent reliability test was conducted to observe the impact of this action. The revised Cronbach's alpha value was obtained as .831.

Table 4.

Initial results of the Cronbach's alpha reliability tests for the awareness scale

	Scale Mean If Item	Corrected Item-Total Correlation	Cronbach's Alpha If Item

Item No.	Washback Scale	Deleted		Deleted
1.	I know what ÖABT (Öğretmenlik Alan Bilgisi Testi) is.	17,80	,178	,807
6.	A separate preparation program (e.g., an online or face-to-face course) beside the university courses is necessary to succeed in the exam.	17,29	,220	,812
7.	ÖABT is a test designed to assess and appoint EFL teachers to public schools, specifically the Ministry of National Education (MEB) schools in Türkiye.	17,42	,184	,812

With a closer look at the results from the Cronbach's Alpha test for the perception and washback items in the scale, it was discovered that the scale exhibited a high level of reliability, with a coefficient of .919. Nevertheless, there were precisely six items that, if removed from the scale, would enhance its reliability (see Table 5). Upon closer scrutiny of the recommended questions, it was discovered that items 15, 33, and 40 were designed to measure attitude towards TFKT rather than perception, which was not assessed in this study. Consequently, it was determined that these items should be excluded. It was also seen that items 18 and 20 measure the same concept. Item 18 tests whether respondents feel TFKT evaluates remembered knowledge, whereas item 20 assesses whether they believe their higher-order thinking skills are not evaluated in the exam. Consequently, these two items were likewise removed. Item 34 was omitted from the scale since it was determined that the same insight regarding the effectiveness of TFKT in assessing talents as an EFL teacher could be obtained through item 22. After removing the suggested six items, a revised scale was subjected to another Cronbach's alpha test. The test yielded a reliability level of .940, indicating an excellent level of internal consistency. Consequently, the perception, awareness, and washback scale was finalized according to the results yielded by Cronbach's alpha test.

Table 5.

Initial results of the Cronbach's alpha reliability tests for the perception and washback scales

Item No.	Perception Scale	Scale Mean If Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha If Item Deleted
15.	ÖABT is a challenging exam.	81,89	-,069	,922
18.	ÖABT evaluates memorized knowledge.	81,86	-,125	,923
20.	ÖABT does not measure my higher-order thinking skills (e.g. analysis, synthesis, and evaluation).	81,44	-,050	,924
Item No.	**Washback Scale**			
33.	I feel nervous about taking ÖABT.	81,52	-,060	,924
34.	I believe that ÖABT does not effectively assess my abilities as an EFL teacher.	81,70	-,232	,925
40.	I am concerned about the difficulty level of the questions in ÖABT.	81,49	,212	,920

Next, the researchers applied Exploratory Factor Analysis for the awareness items, and perception and washback items in the scale separately, as was done in the reliability test. The reason why the items were analyzed separately was that the participants chose the most suitable option out of "Definitely", "Probably", "Not Sure" and "Definitely Not" in the awareness scale, yet they selected "Strongly Agree", "Agree", "Disagree" and "Strongly Disagree" in the perception and washback scale. Based on the results of this analysis, the researchers first examined the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO), which is a test designed to assess the appropriateness of data for factor analysis (Shrestha, 2021). Guttman (1954) states that The KMO value ranges from 0 to 1, and the values ranging from 0.8 to 1.0 suggest that the sampling is sufficient. KMO scores ranging from 0.7 to 0.79 can be considered average, whereas values ranging from 0.6 to 0.69 are of moderate quality. KMO values below 0.6 indicate insufficient sampling and necessitate remedial action (Guttman, 1954). If the number is below 0.5, the

findings of the factor analysis will definitely not be very appropriate for data analysis (Guttman, 1954). Accordingly, the KMO value of the awareness scale was found to be ,815 which proved that the scale had sufficient validity. To ascertain if the observed variables in a data set are sufficiently intercorrelated to be meaningfully merged into a reduced number of components, researchers also employed Bartlett's Test of Sphericity. In this test, by comparing the correlation matrix to an identity matrix, one may determine if the variables are independent and unsuitable for factor analysis or if they are connected and acceptable for factor analysis. A significance value (p-value) less than 0.05 suggests that conducting a factor analysis on the data set may be beneficial. Consequently, the p-value for the awareness scale was measured as < ,001 which showed availability for factor analysis.

Next, the scale was subjected to principal component analysis because the goal was to look at the data and find the fewest factors that would best describe the whole set of data. Cattell (1966) suggested using graphs to find out how many factors there are. On the vertical axis of a scree plot are the eigenvalue magnitudes, and on the horizontal axis are the eigenvalue numbers. To achieve this, a Scree test was employed. By using this test, one may determine how many components should be extracted at the most before the quantity of unique variation starts to dominate the common variance structure (Cattell, 1966). As a result of this, first, the cumulative percentage of the initial eigenvalue was seen as 70,855%, which was the validity value of the scale. Moreover, it was found that the 9 items in the awareness scale comprised two factors. On further analysis of the results, the first 4 items were found to be related to one factor, whereas the remaining 5 were related to another. A more comprehensive examination followed this result, which led to naming the first four items as assesing the awareness of the construct of TFKT and the remaining five items as assessing the awareness of the role of TFKT. In addition, the awareness scale was determined to have construct validity with a factor load over ,40 for each item (see Table 6).

Table 6.

Summary of factors related to awareness of TFKT

Factors		**Factor Loading**
Item No.	**Component 1: Awareness of the construct of TFKT**	
1.	I have reviewed previous ÖABT questions to familiarize myself with the exam.	,715
2.	I know the number of questions that are typically asked in the exam.	,815
3.	I know the types of items or questions that are commonly asked in the exam.	,884
4.	I am knowledgeable about the topics that are assessed in the exam.	,822
Item No.	**Component 2: Awareness of the role of of TFKT**	
5.	ÖABT includes items to assess my overall English Language Proficiency.	,787
6.	ÖABT assesses my knowledge on English Linguistics.	,897
7.	ÖABT assesses my knowledge on English Literature.	,824
8.	ÖABT assesses my knowledge and skills regarding English Language Teaching Methodology.	,888
9.	ÖABT assesses my knowledge and skills regarding English Language Testing and Assessment.	,904

Finally, the communality values were examined for each item in this scale. These values reflect the shared variance between each variable and all other variables in a component analysis. Communalities precisely quantify the extent to which the underlying components extracted in the study explain the variation in each observable measure (Browne, 1969). Researchers often use communalities to assess the overall quality of the factor analysis results and to decide to if the factors that have been found account for a sufficient amount of the variation. High communalities typically indicate that the factors adequately represent the original variables. In contrast, low communalities (under ,40) might suggest potential issues with the factor structure or the suitability of the variables for the analysis (Browne, 1969). In this scale, the minimum communality value was ,518 and the highest was ,817. Therefore, it can be concluded that the items sufficiently represent the factors assessed (see Table 7).

Table 7.

Communalities for the awareness scale proving the validity of the factors assessed

Item No.	Awareness Scale	Initial	Extraction
1.	I have reviewed previous ÖABT questions to familiarize myself with the exam.	1,000	,518
2.	I know the number of questions that are typically asked in the exam.	1,000	,665
3.	I know the types of items or questions that are commonly asked in the exam.	1,000	,782
4.	I am knowledgeable about the topics that are assessed in the exam.	1,000	,685
5.	ÖABT includes items to assess my overall English Language Proficiency.	1,000	,607
6.	ÖABT assesses my knowledge of English Linguistics.	1,000	,817
7.	ÖABT assesses my knowledge of English Literature.	1,000	,705
8.	ÖABT assesses my knowledge and skills regarding English Language Teaching Methodology.	1,000	,787
9.	ÖABT assesses my knowledge and skills regarding English Language Testing and Assessment.	1,000	,810

Extraction Method: Principle Component Analysis

The results of the principle component analysis conducted on for 10 perception items in the scale showed a meaningful correlation among them with a determinant value of ,008. This proved that this scale had construct validity. Based on the findings of the KMO analysis, the value was seen as ,916 with a p-value of < ,001; therefore, the scale was suitable for factor analysis. Moreover, the cumulative percentage of the initial eigenvalue was seen as 62,558%, which was the validity value of the scale. The researchers continued to check the communalities. As explained above, these values should be over ,40 and the lowest value in the perception scale was ,465 and the highest value was ,832 (Table 8). Moreover, it was found that the 10 items in the perception scale were all related to one factor, which also justifies the researchers' decision to name it as perception scale.

Table 8.

Communalities for the perception scale proving validity of the factors assessed

Item No.	Perception Scale	Initial	Extraction
10.	I feel confident in my own knowledge of the topics assessed in ÖABT.	1,000	,832
11.	I believe ÖABT effectively assesses the knowledge and skills we acquire through our lessons at the university.	1,000	,629
12.	ÖABT is an effective assessment tool for appointing qualified EFL teachers to MEB schools.	1,000	,686
13.	ÖABT effectively evaluates the required knowledge and skills for becoming a successful English teacher.	1,000	,723
14.	ÖABT adequately evaluates my ability to apply theoretical concepts to real classroom situations.	1,000	,530
15.	ÖABT is significant for all EFL teachers.	1,000	,465
16.	The score obtained in ÖABT provides an indication of my potential effectiveness as an English language teacher.	1,000	,610
17.	I consider ÖABT to be a fair assessment.	1,000	,541
18.	By preparing for ÖABT, I can enhance my knowledge and skills essential for the English teaching profession.	1,000	,559
19.	ÖABT serves as an effective exam for assessing my English language proficiency.	1,000	,682

For the final section, the washback items in the scale, the same steps in the analysis were followed one last time. The KMO analysis of the 18 washback items in the scale pointed out a value of ,922 which,

meant an excellent construct validity. The significance level according to Bartlett's Test came out as ,000. A result of 0.000 in the output of Bartlett's Test in SPSS usually suggests that the p-value for the test is quite small, potentially smaller than the level of precision utilized in the report (Pett et al., 2003). The statistical value is effectively 0, indicating compelling evidence to reject the null hypothesis. The null hypothesis of Bartlett's Test of Sphericity asserts that the correlation matrix is an identity matrix, demonstrating that there is no population association between the variables (Pett et al., 2003). Thus, a low p-value (often denoted as 0.000 in statistical software) indicates a strong correlation between the variables and confirms that the correlation matrix is not an identity matrix (Pett et al., 2003). A result of 0.000 in Bartlett's Test output indicates that the variables in the dataset are correlated and appropriate for factor analysis. This result validates the application of factor analysis in uncovering latent components within the data, given that the variables have substantial relationships. Following this, the initial eigenvalues show that under the washback scale there are three factors assessed. In other words, the results supported the researchers' decision to categorize the washback scale into TFKT's washback effect on pre-service EFL teachers, washback effect on EFL teacher education, and washback effect on pre-service EFL teachers' future professional lives. Table 9 presents a more detailed categorization of these factors. In addition, when the total variances were scrutinized, the construct validity percentage of the washback scale was obtained as 62,722%. Since it is over 50%, it can be again concluded that the scale has construct validity.

Table 9.

Summary of the factors assessed in the washback scale

Factors		**Factor Loading**
Item No.	**Component 1: Washback effect on pre-service EFL teachers**	
20.	I believe that ÖABT has positively influenced the teaching of English in Türkiye.	,710
22.	ÖABT provides me with an opportunity to review essential knowledge and skills for an EFL teacher.	,713
23.	I believe ÖABT offers an opportunity to improve the quality of EFL teacher education programs in Türkiye.	,721
27.	I think that I can perform better as an EFL teacher thanks to ÖABT.	,747
28.	ÖABT can be advantageous for EFL teachers in their professional growth.	,833
29.	ÖABT encourages me to engage in continuous professional development as an aspiring EFL teacher.	,964
30.	ÖABT motivates me to improve my English language skills and knowledge.	,788
31.	I feel confident in my ability to perform well in ÖABT.	,511
32.	I am motivated to study and prepare for ÖABT.	,485
34.	I feel positive about the impact of ÖABT on my future teaching practice.	,742
36.	ÖABT encourages me to read articles about my field of study.	,588
37.	Thanks to ÖABT, I have a higher level of motivation to study for my courses at university.	,478
Item	**Component 2: Washback effect on EFL teacher education**	
24.	During their lessons, the lecturers in my university actively guide us in the preparation process for ÖABT.	,869
25.	Preparing for ÖABT positively impacts/impacted my academic performance at university.	,580
26.	My university instructors incorporate previous ÖABT questions into their lectures.	,640
33.	My courses at university prepare me for ÖABT.	,905
35.	My university instructors design their course contents according to ÖABT coverage.	,822
Item	**Component 3: Washback effect on pre-service EFL teachers' future professional lives**	
21.	Since I started university, ÖABT has influenced my study habits.	,544

The researchers, then, maintained their inspection of the communalities. As previously stated, it is expected that these values should surpass,40. The washback scale recorded a minimum value of -409 and a maximum value of -824 (Table 10). In addition, it was determined that each of the ten perception scale items pertained to a single factor, which further supports the researchers' choice to designate it as the perception scale.

Table 10.

Communalities for the washback scale proving validity of the factors assessed

Item	Washback Scale	Initial	Extraction
20.	I believe that ÖABT has positively influenced the teaching of English in Türkiye.	1,000	,629
21.	Since I started university, ÖABT has influenced my study habits.	1,000	,549
22.	ÖABT provides me with an opportunity to review essential knowledge and skills for an EFL teacher.	1,000	,565
23.	I believe ÖABT offers an opportunity to improve the quality of EFL teacher education programs in Türkiye.	1,000	,640
24.	During their lessons, the lecturers in my university actively guide us in the preparation process for ÖABT.	1,000	,688
25.	Preparing for ÖABT positively impacts/impacted my academic performance at university.	1,000	,564
26.	My university instructors incorporate previous ÖABT questions into their lectures.	1,000	,577
27.	I think that I can perform better as an EFL teacher thanks to ÖABT.	1,000	,547
28.	ÖABT can be advantageous for EFL teachers in their professional growth.	1,000	,708
29.	ÖABT encourages me to engage in continuous professional development as an aspiring EFL teacher.	1,000	,824
30.	ÖABT motivates me to improve my English language skills and knowledge.	1,000	,680
31.	I feel confident in my ability to perform well in ÖABT.	1,000	,592
32.	I am motivated to study and prepare for ÖABT.	1,000	,586
33.	My courses at university prepare me for ÖABT.	1,000	,751
34.	I feel positive about the impact of ÖABT on my future teaching practice.	1,000	,746
35.	My university instructors design their course contents according to ÖABT coverage.	1,000	,708
36.	ÖABT encourages me to read articles about my field of study.	1,000	,409
37.	Thanks to ÖABT, I have a higher level of motivation to study for my courses at university.	1,000	,626

Following the reliability analysis and EFA, the researchers then proceeded to apply the revised scale to another sample group of 300 respondents. The new data set was analyzed one more time to see if there were any changes to the reliability of the scale. It was found that the Cronbach alpha value for the awareness scale increased to ,874 from ,831, with no items increasing the reliability if omitted from the scale. Moreover, the alpha value for the perception and washback scale also increased from ,940 to ,954. There were also no items increasing the reliability if deleted. Table 11 summarizes the results of Cronbach's alpha analysis conducted on the new data set collected from the second group.

Table 11.

Reliability Statistics of the Scale Based on the Second Sample Group

Scale	Cronbach's alpha	Cronbach's alpha based on Standardized Items	Number of Items
Awareness	,874	,878	9
Perception and Washback	,954	,954	28

4. Discussion and Conclusion

This study aimed to develop a reliable and valid scale to measure the washback effect of TFKT on pre-service EFL teachers and EFL teacher education programs, and to examine the pre-service EFL teachers' level of awareness and perception of TFKT, as part of a more comprehensive study conducted as a Master's thesis (Ünal Aydın, 2024). The feedback obtained from the experts and the results of the statistical analyses confirmed that the scale can assess all at once. The first statistical analysis applied was Cronbach's alpha test which provides a way to evaluate the consistency of responses to the items in the scale. High internal consistency showed that the items consistently measure the same underlying construct, which is critical for ensuring that the scale accurately measures what it is designed to measure. In consistency with this test, the Exploratory Factor Analysis was applied next, and the underlying constructs found were as follows: awareness of the construct of TFKT and awareness of the role of TFKT; perception on TFKT; washback effect on pre-service EFL teachers, washback effect on EFL teacher education and washback effect on pre-service EFL teachers' future professional development.

It is important to recognize that the process of washback research is complex and requires a careful examination, similar to peeling an onion layer by layer; therefore, researchers need to navigate through the complexities and contextual intricacies to build comprehensive validation evidence for the washback impact on teaching and learning (Alla & Norhaslinda, 2024). Various research examined washback from various angles, addressing topics like whether washback exists, what washback looks like, and what causes washback (Cheng et al., 2004). Based on the existing literature (e.g., Alderson & Hamp-Lyons, 1996; Cheng, 1997; Wall & Alderson, 1993), it can be stated that washback primarily impacts teaching content (Choi, 2008); namely, they have identified teaching content as the most susceptible to change due to testing. It is evident from such research that washback is a complex construct, and based on the results of the present study, it can be concluded that whether TFKT as a high-stakes test influences teaching content in English Language Teaching programs as well as how it influences pre-service EFL teachers can now be measured. It is also important to notice that the developed scale consists of measuring the awareness and perception of EFL teachers related to TFKT. How the exam is perceived and how much is known about it is closely connected to its washback effect. Consequently, this study is significant on both a theoretical and a practical level since it can be utilized to see the situation on a larger scale and to make changes in the test itself if necessary.

Looking from a broader perspective, to the best of the researchers' knowledge, the research conducted on teacher recruitment around the world and in Türkiye do not include any investigation into how teacher learning and teacher education programs are impacted through teacher evaluation and recruitment tests. It can be seen from the instances of studies conducted around the world (Balter & Duncombe, 2008; Darling-Hammond, 2002; Howard et al., 2016; Liu & Johnson, 2006; Ochieng, 2006) that they mainly focus on the student side of the issue, not the teaching side. Also, a standardized assessment tool applied nationwide is not preferred except for Thailand, India and some states of the USA (Please see Literature Review), where we still cannot find studies done on the teacher recruitment method itself and its washback effect.

Thus, this study fills a significant void in the existing body of research by developing an innovative scale that is specifically geared toward measuring the washback effect of TFKT on EFL teacher candidates and EFL teacher education programs, and EFL teacher candidates' awareness and perception of the test. In doing so, it offers a specialized instrument that can evaluate the specific impact that this particular test has on this particular population, and makes a contribution toward improving the precision and accuracy of assessing the washback effect. Moreover, this scale contributes greatly to washback research in the Turkish context by presenting a focused and contextually appropriate instrument. By understanding the various dimensions of washback, teacher educators and policymakers can make informed decisions to improve the design and implementation of such tests. It is also believed that the study has the potential to encourage and facilitate further research on the impact of high-stakes tests on educational outcomes and the professional development of pre-service teachers, not only in the Turkish context but potentially in similar contexts globally because it establishes a reliable and valid measure for assessing the washback effect in this specific context.

In conclusion, this scale development study contributes to the growing body of research on the washback effect in the field of language testing and assessment, particularly in the context of EFL teacher education. By capturing the multifaceted nature of washback, the scale offers insights into the complex

relationship between high-stakes testing and teaching and learning practices. Moving forward, further research is needed to validate the scale across different samples and to explore its potential applications in improving the quality of EFL teacher evaluation and recruitment as well as teacher training programs.

Acknowledment

This article constitutes a component of the Master's thesis entitled “The washback effect of English Teacher Field Knowledge Test on pre-service EFL teachers and EFL teacher educators and their level of awareness and perception in Türkiye: A scale development study”, submitted in partial fulfillment of the requirements for the degree of Master of Arts in English Language Teaching Program, at Başkent University. We, as the authors of this manuscript, acknowledge that there is no conflict of interest in this study.

Note on Ethical Issues

The authors confirm that ethical approval was obtained from Başkent University (Approval Date: 30 /01 /2023).

References

Akpınar, K. D., & Çakıldere, B. (2013). Washback effects of high-stakes language tests of Turkey (KPDS and ÜDS) on productive and receptive skills of academic personnel. *Journal of Language and Linguistic Studies, 9*(2), 81-94.

Alderson, J. C. (2004). ‘Foreword’ in L. Cheng, Y. Watanabe, and A. Curtis (eds.). Washback in language testing: Research contexts and methods. London: Lawrence Erlbaum.

Alderson, J. C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback. *Language testing, 13*(3), 280-297.

Alla Khan & Norhaslinda Hassan (2024). Washback into an Ecosystem of Teaching, Learning and Testing within Asia and Beyond: An Interview with Liying Cheng, *Language Assessment Quarterly, 21*(1), 100-112. https://doi.org/10.1080/15434303.2023.2276917

Allen, M. B. (2005). Eight Questions on Teacher Recruitment and Retention: What does the Research Say?. *Education Commission of the States (NJ3)*.

Andrews, S., Fullilove, J., & Wong, Y. (2002). Targeting washback-a case-study. *System, 30*(2), 207-223.

Atav, E., & Sönmez, S. (2013). Öğretmen adaylarının kamu personeli seçme sınavı (KPSS)’na ilişkin görüşleri. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 2013*(1), 1-13.

Audet, L. A., Lavoie-Tremblay, M., Tchouaket, É., & Kilpatrick, K. (2023). The level of adherence to best-practice guidelines by interprofessional teams with and without acute care nurse practitioners in cardiac surgery: A study protocol. *Plos one, 18*(3). https://doi.org/10.1371/journal.pone.0282467

Balter, D., & Duncombe, W. D. (2008). Recruiting highly qualified teachers: do district recruitment practices matter?. *Public Finance Review, 36*(1), 33-62.

Biggs, J. (1999). What the student does: Teaching for enhanced learning. *Higher education research & development, 18*(1), 57-75.

Browne, M. W. (1969). Fitting the factor analysis model. *Psychometrika, 34*(3), 375-394.

Burrows, J. (1999). Going beyond labels: A framework for profiling institutional stakeholders. *Contemporary Education, 70*(4), 5.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate behavioral research, 1*(2), 245-276.

Cheng, L. (1999). Changing assessment: Washback on teacher perceptions and actions. *Teaching and teacher education, 15*(3), 253-271.

Cheng, L. (2005). *Changing language teaching through language testing: A washback study* (Vol. 21). Cambridge University Press.

Cheng, L. (1997). How does washback influence teaching? Implications for Hong Kong. *Language and Education, 11*, 38-54.

Cheng, L., & Curtis, A. (2004). Washback or Backwash: A Review of the Impact of Testing on Teaching and Learning. In L. Cheng, Y. Watanabe, & A. Curtis (Eds), Washback in Language Testing (pp. 8). Lawrence Erlbaum Associates, Inc.

Cheng, L., Watanabe, Y. and Curtis, A. (Eds). (2004). *Washback in language testing: Research contexts and methods*. Mahwah, NJ: Lawrence Erlbaum.

Choi, I. (2008). The impact of EFL testing on EFL education in Korea. *Language Testing, 25*(1), 39-62.

Cohen, R. J., Swerdlik, M. E., & Phillips, S. M. (1996). *Psychological testing and assessment: An introduction to tests and measurement*. Mayfield Publishing Co.

Creswell, J. W. (2009). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* (3rd ed.). Sage Publication, California.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston, Orlando.

Çakır, İ. (2017). The washback effects of secondary education placement examination on teachers, school administrators and parents with specific reference to teaching English as a foreign language. *Turkish Journal of Teacher Education, 6*(2), 61-73.

Darling-Hammond, L. (2002). The research and rhetoric on teacher certification: A response to "teacher certification reconsidered". *Education Policy Analysis Archives, 10*(36), 1-55.

DeVellis, R. F., & Thorpe, C. T. (2021). *Scale development: Theory and applications*. Sage Publications.

Elmaci, S. (2015). Kamu personeli seçme sınavı ve alan bilgisi sınavına ilişkin öğretmen görüşlerinin ve metaforik algılarının belirlenmesi (Master's thesis, Sosyal Bilimler Enstitüsü).

Erdoğan, P. (2019). A Study on pre-service efl teachers' admission into teaching programs and EFL teachers' recruitment in Turkish context.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American psychologist, 39*(3), 193.

Goldhaber, D., & Hansen, M. (2010). Race, gender, and teacher testing: How informative a tool is teacher licensure testing?. *American Educational Research Journal, 47*(1), 218-251.

Guttman, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika, 19*(2), 149-161.

Hatipoğlu, Ç. (2010). Summative Evolution of an Undergraduate 'English Language Testing and Evaluation' Course by Future English Language Teachers. *English Language Teacher Education and Development (ELTED), 13*, 40-51.

Hatipoğlu, Ç. (2016). The impact of the university entrance exam on EFL education in Turkey: Pre-service English language teachers' perspective. *Procedia-Social and Behavioral Sciences, 232*, 136-144.

Hawkey, R. (2006). *Impact theory and practice: Studies of the IELTS test and Progetto Lingue 2000*. Cambridge: Cambridge University Press.

Howard, A., Basurto-Santos, N. M., Gimenez, T., Moncada, A. M. G., McMurray, M., & Traish, A. (2016). A comparative study of English language teacher recruitment, in-service education and retention in Latin America and the Middle East. Publication City, Country: Publisher.

Hughes, R. (2004). Testing the visible: literate biases in oral language testing. *Journal of Applied Linguistics, 1*(3).

Johnstone, R. (2004). Language teacher education. In A. Davies & C. Elder (eds.), *The handbook of applied linguistics* (pp. 649-671). Malden, MA: Blackwell Publishing Ltd. http://dx.doi.org/10.1002/9780470757000.ch26

Karaer, H., Karaer, F., & Kartal, E. (2018). Opinions of teacher candidates towards the teaching field knowledge tests on the public personnel selection examination. *Erciyes Journal of Education, 2*(2), 40-58.

Karataş, T. Ö., & Okan, Z. (2021). The Powerful Use of an English Language Teacher Recruitment Exam in the Turkish Context: An Interactive Qualitative Case Study. *International Online Journal of Education and Teaching, 8*(3), 1649-1677.

Kılıçkaya, F. (2016). Washback effects of a high-stakes exam on lower secondary school English teachers' practices in the classroom. *Lublin Studies in Modern Languages and Literature, 40*(1), 116-134.

Külekçi, E. (2016). A concise analysis of the Foreign Language Examination (YDS) in Turkey and its possible washback effects. *International Online Journal of Education and Teaching, 3*(4), 303-315.

Liu, E., & Johnson, S. M. (2006). New teachers' experiences of hiring: Late, rushed, and information-poor. *Educational Administration Quarterly, 42*(3), 324-360.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*(3), 241-256. https://doi.org/10.1177/026553229601300302

MoNE (2015). Milli Eğitim Bakanlığı Öğretmen Atama ve Yer Değiştirme Yönetmeliği. TC. Resmi Gazete, 29329, 17 Nisan 2015. https://www.resmigazete.gov.tr/eskiler/2015/04/20150417-4.htm Erişim: 27.01.2021

Moser, K. (2012). Does Praxis make perfect? A personal journey through the Praxis II: World Language Test. The Clearing House: *A Journal of Educational Strategies, Issues and Ideas, 85*(4), 123-128.

Nettles, M. T., Scatton, L. H., Steinberg, J. H., & Tyler, L. L. (2011). Performance and passing rate differences of African American and white prospective teachers on PraxisTM examinations: a joint project of the National Education Association (NEA) and Educational Testing Service (ETS). *ETS Research Report Series, 2011*(1), i-82.

Ochieng, C. (2013). Classroom-based factors influencing teaching and learning in Public Primary Schools in Ukwala Division of Siaya County, Kenya (Doctoral dissertation, University of Nairobi).

Özmen, K. S. (2011a). Analysing washback effect of SEPPPO on prospective English teachers. *The Journal of Language and Linguistic Studies, 7*(2), 24-52.

Özmen, K. S. (2011b). Washback effects of the inter-university foreign language examination on foreign language competencies of candidate academics. *Novitas-ROYAL (Research on Youth and Language), 5*(2), 215-228.

Pearson, I. (1988). Tests as levers for change. In D. Chamberlain & R.J. Baumgardner (Eds.), *ESP in the classroom: Practice and evaluation* (pp. 98-107). Modern English, London.

Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research.* Sage.

Pituch, K. A., & Stevens, J. P. (2016). *Applied Multivariate Statistics for the Social Sciences*. 6th ed. New York and London.

Sadeghi, K., Ballıdağ, A., & Mede, E. (2021). The washback effect of TOEFL iBT and a local English Proficiency Exam on students' motivation, autonomy and language learning strategies. *Heliyon, 7*(10).

Sayın, B. A., & Aslan, M. M. (2016). The negative effects of undergraduate placement examination of English (LYS-5) on ELT students in Turkey. *Participatory Educational Research, 3*(1), 30-39.

Scaramucci, M. V. (2002). Entrance examinations and TEFL in Brazil: a case study. *Revista Brasileira de Linguística Aplicada, 2*, 1-13.

Seçer, İ. (2018). *Psikolojik test geliştirme ve uyarlama süreci: SPSS ve LISREL uygulamaları*. Anı yayıncılık.

Sert, C. (2015). The role of teacher field knowledge test on teachers' knowledge. *Procedia-Social and Behavioral Sciences, 199*, 801-805.

Aktuğ, C. S. (2018). An evaluation of teacher domain-specific knowledge test for teacher candidates of English in Turkey.

Sevimli, S. (2007). *The washback effect of foreign language foreign language component of the university entrance examination on the teaching and learning context of English language groups in secondary education (A case study).* [Unpublished MA Thesis, Gaziantep University].

Shohamy, E. (1993). The Power of Tests: The Impact of Language Tests on Teaching and Learning. NFLC Occasional Papers.

Shohamy, E. (2017). The discourse of language testing as a tool for shaping national, global, and transnational identities. In *The Discourse of Culture and Identity in National and Transnational Contexts* (pp. 115-126). Routledge.

Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: Washback effect over time. *Language Testing, 13*(3), 298–317. https://doi.org/10.1177/026553229601300305

Shrestha, N. (2021). Factor analysis as a tool for survey analysis. *American Journal of Applied Mathematics and Statistics, 9*(1), 4-11.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*(1), 1-22.

Şeker, H., & Gençdoğan, B. (2006). *Psikolojide ve eğitimde ölçme aracı geliştirme*. Nobel.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International journal of medical education, 2*, 53.

Taylor, L. (2005). Washback and impact. *ELT Journal, 59*(2), 154-155.

Taylor, R., Hardman, M., Riordan, S., Pillinger, C., & Moss, G. (2023). *Teacher quality, recruitment, and retention: Rapid Evidence Assessment*.

Thomas, S., Smees, R., Madaus, G. F., & Raczek, A. E. (1998). Comparing Teacher Assessment and Standard Task Results in England: the relationship between pupil characteristics and attainment. *Assessment in Education: Principles, Policy & Practice, 5*(2), 213-246.

Thornbury, S. (1997). *About Language: Tasks for Teachers of English*. Cambridge University Press.

Wall, D. & Alderson, J. (1993). Examining washback: The Sri Lankan impact study. *Language Testing, 10*, 41-69.

Watanabe, Y. (1996). Investigating washback in Japanese EFL classrooms: Problems of methodology. *Australian review of applied linguistics. Series S, 13*(1), 208-239.

Verma, J. P., & Abdel-Salam, A. S. G. (2019). *Testing statistical assumptions in research*. John Wiley & Sons.

Yağcı, E., & Kurşunlu, E. (2017). Öğretmen adaylarının aldıkları mesleki eğitimin yeterliliğine ve Kamu Personeli Seçme Sınavı'na (KPSS) yönelik görüşlerinin incelenmesi. *Uluslararası Eğitim Programları ve Öğretim Çalışmaları Dergisi, 7*(14), 1-14.

Yeşilçınar, S. (2018). *An Evaluation of EFL Teacher Assessment and Evaluation: A Suggested Model*. [Doctoral Dissertation, Gazi University Institute of Educational Sciences]. Ulusal Tez Merkezi. https://tez.yok.gov.tr/UlusalTezMerkezi/TezGoster?key=fS4sqEZr79C_n60Rk6MjFX0kpOrNW5yNkjJWMTkQfWmsbZMWgRAYHAsd97d3kpl1

Yeşilçınar, S., & Çakır, A. (2020). Development and validation of the English teachers' attitudes towards recruitment system scale. *Ilkogretim Online, 19*(3).

Yıldırım, Ö. (2010). Washback effects of a high-stakes university entrance exam: Effects of the English section of the university entrance exam on future language teachers in Turkey. *The Asian EFL Journal Quarterly, 12*(2), 92-116.