International Journal of Contemporary Educational Research (IJCER)

www.ijcer.net

Mapping Teacher-Produced Tests to a Usefulness Model

Cemile Doğan[1]
[1]Necmettin Erbakan University, 0000-0002-5246-6692

Article History
Received: 25.04.2023
Received in revised form: 20.08.2023
Accepted: 01.09.2023
Article Type: Research Article

To cite this article:

Doğan, C. (2023). Mapping teacher-produced tests to a usefullness model. *International Journal of Contemporary Educational Research*, *10*(3), 635-648. https://doi.org/10.52380/ijcer.2023.10.3.456

International Journal of Contemporary Educational Research
Volume 10, Number 3, September 2023, Page 635-648 ISSN: 2148-3868

Mapping Teacher-Produced Tests to a Usefulness Model

Cemile Doğan[1*]
[1]Necmettin Erbakan University

Abstract

Tests are designed as an integral part of the teaching process, necessarily including stakeholders from the onset of preparations to grade allocation, the administration of the test, and the interpretation of the results. The process commences with selecting content to evaluate, deciding upon the skills to be tested, and meeting course objectives (Giraldo & Murcia Quintero, 2019; O'Louglin, 2013; Vogt and Tsagari, 2014). Several questions arise in terms of how to standardize the development process and evaluate its usefulness. Typically, what is the best test for our context? What does this test actually test? What relevant information does the test provide? How does this test affect teaching and learning behavior? And in what ways is the test useful? Although each language program's particular needs may differ, the answers given to the questions above provide a basis for institutional decisions. None are set in stone, and at their root is the critical role testing plays in facilitating what gets learned. The current study initiated action to develop and analyze an achievement test specifically designed for a compulsory A1-level English course delivered to all freshmen students enrolled in Turkish-medium departments at state universities across Türkiye. 150 students who are enrolled in several undergraduate programs at the Faculty of Education at a state university constituted the universe of the study. The researcher analyzed the test after administration and mapped the qualities according to a test usefulness model, aiming to address the research gap regarding quality teacher-produced tests.

Keywords: ELT, Test Usefulness, Measurement and Evaluation, Teacher-Produced Tests

Introduction

Language tests are a component of the teaching and learning process, and they are used for a variety of purposes that are mainly divided into two broad categories. The first use of language tests is to make inferences about the language abilities of the students, that is, the extent to which they can use language to perform tasks in life (Bachman, 1991, p. 680). It can be inferred from the statement that a test has the purpose of measuring the test takers' ability or competence level in a given domain and their capacity for non-test language use. The second main use of language tests is to refer to them when making decisions based on what is inferred from test scores (Ballıdağ & Karagül, 2021; Bachman & Palmer, 1996; Coombs et al., 2018; Heaton, 1990; Hughes, 2010; Luoma, 2001). The decisions are for selection, diagnosis, placement, progress, grading, certification, and employment purposes. These two major roles indicate the significance of using a concrete, valid framework in designing, implementing, and scoring. A sound test enables us to make inferences about levels or profiles of ability, or predictions about the capacity to perform future tasks in non-test language use contexts. Bachman (1991) argues that the language abilities measured by testing should correspond to the features of a target language's context, depending on the settings. In an instructional setting, a test can be designed to measure the degree of the learners' mastery of the language skills covered in the curriculum. It is essential that the content of the test be representative of the content of the course. The correspondence between the course content and the test provides grounds for interpreting test scores. In other words, test scores confirm what the students and teachers have mastered. There is limited research concerning teacher-produced tests, as they may not lend themselves to as much introspection as standardized, high-stakes tests. Thus, the measurement and evaluation skills of teachers and how their skills relate to classroom instruction are worth thorough analysis to abolish the discrepancy between evaluation practices in various contexts (Galluzzo, 2005; Marzano, 2020; Mertler, 2004; Shohamy, 2020). Inservice professional development programs can provide teachers with reflective opportunities to collaborate on practices to better their testing development and take essential test qualities into consideration (Bonner et al, 2018; Cizek, 2000; Fulcher, 2012; Höl, 2023; Shohamy, 2020). Teachers can

* Corresponding Author: *Cemile Doğan, cemiledogan370@gmail.com*

participate in test preparation, administration, and scoring processes through which there is the possibility of positive backwash for the accountability of classroom instruction, tests, and schools (Bachman, 2000; Brown, 2013; Cochran, McCallum, & Bell, 2010; Sahlberg, 2006).

Literature Review

An overview of the history of test development theory and practice would shed light on the theoretical background of the present study and how it influenced its preparation and analysis. In his seminal article 'Linguistics and Language Testers,' Spolsky divides up the developments in approaches to testing a second language into three distinct eras (Spolsky, 1978). In what he calls the pre-scientific era, tests were based on the expertise of the test writers, which would, at the time, suffice for a test to be considered valid and reliable. This period was followed by a 'psychometric structuralist' approach era in which one could observe the effects of general trends in approaching social sciences. The period, thus, perceived language to be dissected into its components as isolated sounds, words, or structures without a context. It relied heavily on the comparison of L1 and L2 and assumed that the difficulties in second language learning depended on the differences between L1 and L2 (Brown, 2013; Lado, 1964; Luoma, 2001; Willis, 2003). At the time, statistical analysis of items in classical tests was done (Stansfield, 2008), but not intensively. In the 1970s, the 'integrative-sociolinguistic' approach was widely accepted, and statistics was widely used for the analysis of the tests, going beyond merely item analysis (Bachman, 2000; Huot, 1990; Lantolf & Frawley, 1985; Oller, 1976). Oller found via statistical analysis that compartmentalizing language was not a reliable path to assess language proficiency. Cloze tests, which at the time became widely used and remained popular until today, were very effective for testing grammar in context (Abraham & Chapel, 1992; Brown, 2013; Darwesh, 2010; Spolsky, 2002). The debate about the validity and reliability of tests as a whole in representing language proficiency led to the need to define language proficiency. The debate was also triggered by the introduction of the communicative approach to language teaching by Hymes (1972) and Halliday (1973). In the 1980s, Canale and Swain began publishing the Applied Linguistics Journal, with which they legitimized the importance attached to the teaching of 'communicative competence,' encompassing grammatical, sociolinguistic, and strategic competence, and the testing of these competences (Canale & Swain, 1980; Read & Chapelle, 2001; Kirschner, Spector-Cohen, & Wexler, 1996). In the 1990s, language testing specialists felt the need to discuss common professional and ethical aspects of language testing, which led to the birth of the International Language Testing Association (ILTA), whose official publications were Language Testing and Language Testing Update. ILTA was followed by the Association of Language Testers in Europe (ALTE) in 1990 and the European Association for Language Testing and Assessment (EALTA) (Wu & Stansfield, 2001). The foundation of many testing and assessment-oriented associations on a national and international scale worldwide has made language testing and research more professional and collaborative.

The 21st century has boosted creativity in research methodology, factors that affect performance, authentic assessment concerns, and the ethics of language testing (Bachman, 2000). Testing is rooted in classical theory that has been extended, as in Generalizability Theory which helps testers understand the effects of the sources of measurement errors (for an overview of G-theory, see Bachman, 1997). Item response theory (IRT) has become a widely used tool as a measurement model that allows for statistical properties of items and abilities of test takers in large-scale language proficiency tests (Bachman & Eignor, 1997; Pollitt, 1997). The Rasch model is still commonly used in language testing (Adams et al., 1987; Lynch et al., 1988; McNamara, 1991; Bolt, 1992) to investigate the effects of raters and tasks in language performance assessments (Brown, 1995; Lumley & McNamara, 1995; Weigle, 1998). Test takers' background features, their use of strategies, and the relationship between language aptitude, proficiency, and intelligence have been intense topics in Structural Equation Modelling (SEM) (Ginther and Stevens, 1998; Purpura, 1997; Sasaki, 1996). Recently, however, the general tendency in trait perspectives has been to integrate quantitative approaches such as G-theory, IRT, and SEM into an analytic paradigm.

Qualitative research approaches have also become widespread as they are used to include test takers' characteristics, processes, and tactics, as well as the description of the discourse created in language assessments (Banerjee & Luoma, 1997; Cochran, McCallum, & Bell, 2010; Fulcher, 2012; Horwitz, 2001; Latif & Wasim, 2022; Shohamy, 2001). Retrospective and introspective verbal reports, observations, questionnaires, interviews, and discourse analysis are within the scope of qualitative approaches to testing. As language testers become more experienced in combining quantitative and qualitative measures, the results of language tests are reported in both fashions (Clapham, 1996; Sasaki, 1996). This eases the focus on traditional areas of linguistic competence and the four main skills of language, which lead to the development of prototype test instruments. These test instruments embody a selection of task types, such as multiple choice, response items with cues, structured interviews, and self-assessment checklists, which are particularly relevant in the communicative framework of language testing. There has been a tendency towards 'performance' assessment and a movement towards what has been referred to variously as 'alternative assessment' (Aschbacher, 1991; Terwilliger, 1998). 'Authenticity' as a relative quality, has become one of the qualities of a good test and has displayed itself as

integrated skills items (Lewkowicz, 2000; Yamtim & Wongwanich, 2014). Validity, reliability, interactiveness, practicality, and impact proceed from authenticity (Alderson and Hamp-Lyons, 1996; Bailey, 1996; Wall, 1996; Shohamy et al., 1997; Cheng, 1999). As language testers have become more concerned about ethical issues, they report more on their own test development experiences and dig deeper into the issue of professionalism in conducting ethical tests.

Parallel to the shift from the 20th century's discrete outlook towards a more standardized frame with context-specific implementations in language testing, test qualities have become essential in Türkiye as well. A standardized frame with context-specific implementations bears significance because 'in Türkiye, English is currently the only foreign language that has become a compulsory subject at all levels of education' (Kırkgöz, 2008, p. 667). Language testing receives substantial attention from stakeholders. They are officials, field authorities, educational administrators, teacher educators, and teachers. In a higher education context, in a similar vein, preparing quality language tests is demanding. The Basic English Course is one of the compulsory courses in the Higher Education Council curriculum for Turkish-medium universities. However, there is an unequal distribution of social, cultural, technical, and educational opportunities for language instruction and testing (Alan, 2003; Atay, 2008; Ballıdağ, 2020; Ekşi, 2010; Gültekin, 2007; Şentuna, 2002; Tomak & Karaman, 2013). Commercial tests, which are designed by test experts, are insufficient for addressing contextual needs. Therefore, teachers who are delivering the courses hold the most important role in classroom tests. Moreover, because the Basic English Course is mandatory for all students, there is a need to standardize both content and quality and refocus on the course objectives comprehensively.

The course objectives in the current study's context are determined depending on the objectives of the annual syllabus, and class materials and tests are prepared accordingly. Although teacher-produced tests have been regarded skeptically concerning their quality, and test items on standardized tests are generally written by test specialists, pretested, and refined, teachers' awareness can be raised, and through classroom practice research case studies, the quality of the teacher-prepared achievement tests can be increased. Hence, as teachers' awareness is raised through taking responsibility for their classroom tests, they may start with determining the purpose of their tests, from checking the rate of progress to diagnosing existing and/or probable weaknesses regarding the teaching and testing process. The formative tests may serve many purposes; one of the most important of these is balancing teaching according to the assessment technique (Ballıdağ, 2020; Fanrong & Bin, 2022; Llosa, 2011). Another benefit is determining the test techniques that are appropriate for the students' needs in target language use. It is an effective tool to bring contextual needs and teaching techniques together. Summative assessment, on the other hand, is used to assess student performance at the end of the instructional period in the form of end-of-unit or semester exams. At university, midterms, quizzes, and assignments can be considered formative testing; final exams can set an example of a summative test, by the score of which students' success or failure in the course is determined regarding a semester in the academic year.

What is necessary in educational settings with a diverse teaching staff who come from different educational backgrounds and have various approaches to language teaching is to agree on a fundamental understanding of the nature of classroom instruction and create detailed test specifications to make a match between test specifications and test items. As the next step, different test method characteristics to be used are discussed depending on the level, length, syntactic complexity, type of response required, participants, purpose, and topical content. In the light of these data, teachers can produce their tests, which may be more valid and relevant to students' contextual needs as long as they are well constructed. Although the quality of teacher-produced tests has been a topic of concern among educators, who claim that they are low in quality compared to standard tests written by test specialists, Coniam (2009) states that, given the right conditions of time and support, it is possible to produce better-quality test items for practicing teachers as well and sustain a professional, reliable, and valid test.

The current study aimed to develop an achievement test for an A1-level Basic English course and evaluate the qualities of test usefulness through the stages of test specification, moderation, and item analysis. Additionally, it set out to initiate action to create 'good' tests for institutional use and form a test bank for the compulsory course delivered to freshmen students.

Research Design

It is not too much to claim that every stage of test construction requires conscientious work, as scores play a crucial role in the future success of the learners. The course objectives should be defined clearly and applied effectively in the classroom. As a result, tests may have 'a positive washback on the content of classroom teaching' (Fulcher, 2012; Lan & Fan, 2019; O'Loughlin, 2013; Sariyildiz, 2018). The Basic English Course serves many freshmen students at Turkish-medium universities. The researcher set out to formulate a unified set of guidelines for developing and evaluating language tests in light of the literature, which suggests that the adoption of such guidelines has a positive effect on classroom instruction. (Fulcher, 2012; Krashen, 1982; Latif & Wasim, 2022; Hughes, 2010; Ölmezer-Öztürk & Aydin, 2019). The study was conducted at the Faculty of Education in a state university where the medium of instruction is Turkish, and two-hour Basic English Courses

are compulsory according to the curriculum designed by the Higher Education Council that is applicable for use in all Faculty of Education departments in Türkiye. The courses are delivered not only by English teachers but also by departmental academic staff who have previously achieved a score between 60 and 100 in the Centralized Foreign Language Test (YDS) and/or Higher Education Council Language Test (YÖKDİL), both of which are standardized exams administered regularly by the Student Selection and Placement Center in Türkiye. This affords a variety of language instruction provided by academics who may not necessarily have English language teaching experience. They may also possess different beliefs about language teaching and testing in students from different disciplines (Balbay et al., 2018). It is not possible to develop an achievement test that would meet the expectations of the various programs on offer at a typical faculty; however, providing an example that is based on clearly defined stages and presenting it for common use may serve to enhance test design in similar educational settings. The following developmental steps took place in test preparation:

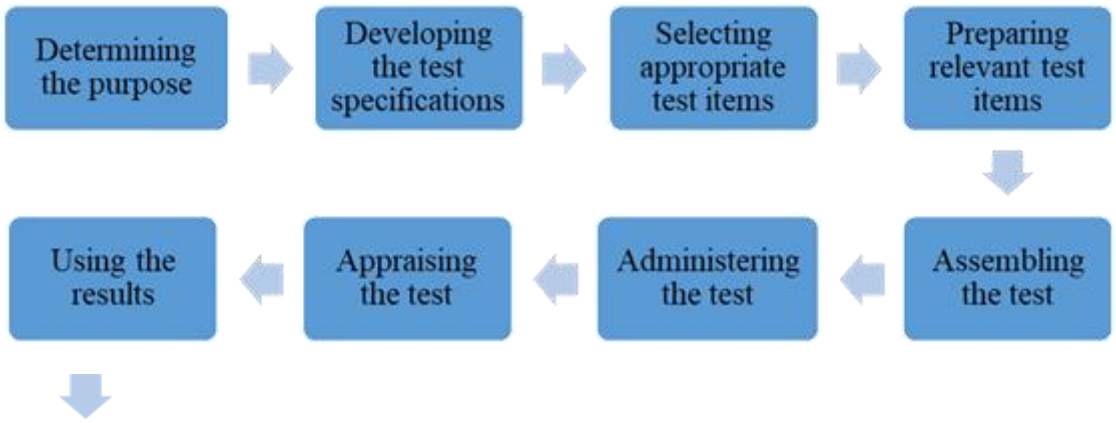

Figure 1. Steps in test preparation (Grondlund & Linn, 1990, p. 110)

Method

Purpose of the test

The purpose of this test is to measure students' knowledge and use of specific lexical and grammatical forms and the degree of their control in comprehending reading passages and writing a paragraph suitable to their level. The results obtained from the test were intended to be used for several reasons, one of which is making decisions about students' progress at the end of the term and the degree to which they have mastered the course objectives. The principles for developing the test were based on Bachman and Palmer's Test Development Suggestions (see Appendix 1). Furthermore, this test can be used to diagnose strengths and weaknesses and decide what aspects of the course might require review to assign students remedial activities.

Participants and Sampling Procedures

The students were from different undergraduate teaching programs at the Faculty of Education, such as Turkish Language, Turkish Language and Literature, Preschool, Geography, and Primary School Teaching. They had been admitted to the faculty based on their national, centralized, and standardized university placement exam. Although the total population of students is nearly 300, including students repeating the course, 150 students (40 males and 110 females) were chosen randomly for the study. Their ages ranged between 18 and 21. They were mainly Turkish students, mostly from the Central Anatolian part of Türkiye with similar educational backgrounds (Turkish state high school graduates). They took two 45-minute Basic English courses per week. A beginner-level course book (Campus Life) prepared according to the guidelines of the CEFR and its grammar component were used with supplementary materials from the internet and other sources if need be. In terms of language proficiency, they are beginner-level students who scored less than 60 (out of 100) in the English proficiency exam officially administered to students of all faculties by the School of Foreign Languages at the beginning of every academic year. Students who score 60 or more in the exam do not have to take the Basic English Course. Students who score 59 or less have to take Basic English Courses I and II during

their freshman year. For diagnostic purposes, the scores of the participants were reviewed. They ranged between 15-45. They reported that they had no exposure to English at all except for social media outside of class.

Descriptions of the Target Language Use Domain and Tasks

The overall aim of the course in the faculty curriculum is to develop basic language skills so that students can follow academic-level resources related to their areas and express themselves in academic environments. The students take the Basic English Course for two semesters in their first year, Basic English I and II, and do not receive further English language training at university. Concerning the achievement test developed, it is aimed at providing feedback on how well the students have learned the content of the course. So, the tasks were based on the language instructional domain. However, considerable effort was put into making a match between the characteristics of the instructional tasks to make them relevant to the real-life domain to maximize authenticity.

Definition of Constructs to be Measured

For this achievement test, a course book syllabus-based construct definition is used. This involves students' ability to recognize and produce correct forms of language (grammar), their ability to read passages appropriate to their level, their ability to provide suitable responses to the given situations (functional language), their ability to write short paragraphs following the provided guidelines, and their appropriate vocabulary use. The specific components below are included in the construct definition:

Research design and operationalization

At the operationalization stage, test specifications (see Appendix 2) were prepared for the various types of tasks that would be completed during the examination. Afterwards, actual test tasks were prepared, instructions were written, and procedures to be followed for scoring were determined.

Developing Test Specifications

While developing test tasks, the following table was used as a basis to prepare a content-valid test for the course.

Table 1. The distribution of skills and areas to be tested and objectives

Objectives	Reading	Grammar	Vocabulary	Functional Language	Writing	Total (%)
Infer the word's meaning from the context of the passage	5					5
Find appropriate answers to questions answered explicitly or in paraphrase	2					2
Understanding relationships among ideas in a text (cause-effect)	2					2
Find specific details (any surface-level information)	1					1
Understand explicitly stated ideas	6					6
Identify the third person's simple verb form		1				1
Discriminate the use of simple negative forms for different personal pronouns		1				1
Supply the correct preposition of time		3				3

Supply the correct preposition of place		1				1
Locate the necessary form of -be for there is/are and that/this/these/those expressions		2				2
Choose the correct expression of quantity		3				3
Supply the correct article		1				1
Recognize the correct use of possessive pronouns		1				1
Distinguish the use of can for ability		1				1
Identify cohesive devices and their functions in a text		1				1
Recall the relevant vocabulary			4			4
Write a paragraph about where he or she lives					1	1
Provide appropriate responses to the given situations				5		5
Total (%)	16	15	5	5	1	41

Inventory of available resources and plan for their allocation

The test was prepared by the researcher. Invigilators were research assistants in the departments. The tests were scored by the researcher and the other two teachers delivering the same course. Developing, administering, and scoring the tests was a normal teaching requirement for the course teachers. Since the number of students taking the course was quite high, scoring took a long time and required considerable attention. The test was administered in students' regular classrooms on a specified date in the academic calendar. It consisted of four pages. Students were expected to give their responses on the test papers. No additional sheet was provided.

Test Structure

The test is divided into four parts: It is designed as a 41 -item test that is made up of:

1. a 15-item reading comprehension, six of which consist of true or false statements
2. a 20-item multiple-choice completion test, including 4 items of vocabulary
3. 1 guided paragraph-writing item
4. a 5-item multiple-choice dialogue completion

The topics chosen are relevant to the topics in the course book. The allotted time for the test is 60 minutes.

Scoring Procedures

All responses were scored first by the researcher and two other teachers giving the same course. Apart from the writing part, the items were scored objectively according to a scoring key. The writing part of the exam was scored according to a writing rubric specified by CEFR.

Administration and analysis of the test

The test was prepared by the researchers. After the test was administered, the papers were scored. The test items were analyzed for the identification of item difficulty and item discrimination, which contribute to reliability measures. The following test item analysis table was formed according to the results:

*Table 2***.** Item Analysis

Item	Item	Item	Item	Item	Item

number	difficulty	discrimination	number	difficulty	discrimination
1*	0.22	0.26	21	0.36	0.42
2	0.78	0.47	22	0.61	0.38
3	0.69	0.32	23	0.81	0.93
4	0.57	0.49	24	0.79	0.81
5	0.26	0.32	25	0.78	0.93
6	0.62	0.30	26	0.87	0.71
7	0.78	0.89	27	0.42	0.58
8	0.74	0.69	28	0.39	0.54
9	0.52	0.36	29	0.25	0.31
10	0.40	0.33	30	0.28	0.32
11	0.61	0.55	31	0.20	0.33
12	0.56	0.42	32	0.67	0.57
13	0.63	0.30	33	0.62	0.70
14	0.30	0.29	34	0.41	0.30
15	0.34	0.30	35	0.53	0.72
16*	0.42	0.27	36*	0.37	0.15
17	0.74	0.91	37	0.22	0.31
18	0.57	0.54	38	0.47	0.45
19	0.30	0.32	39	0.36	0.39
20	0.36	0.47	40	0.44	0.56

The average of correct answers	20.651
The average of item difficulty	0.507
Standard deviation	5.178
KR	0.697

Table 3. Item discrimination criteria

0.40 and above	Fairly discriminating
0.39 - 0.29	Discriminating
0.29 - 0.19	Needs improvement
0.19 and below	Needs to be replaced

questions *1 (0.26), * 16 (0.27) and* 36 (0.15)

The average index of difficulty was 0.507, which was satisfactory. The items with a difficulty level of 70 or above were accepted as easy. 9 questions (2, 7, 8, 17, 23, 24, 25, 26, and 35) were answered correctly by most of the students. The questions with a difficulty level of 30 or below were accepted as difficult items. 8 questions (1, 5, 14, 19, 29, 30, 31, and 37) were difficult questions. The facility value of the remaining items (23 questions) varied within the range of 0.31 and 0.69. In general, the items in the test have a desirable level of difficulty.

Results Regarding the Usefulness of the Test

Certain types of reliability, validity, and practicality are emphasized in Bell's (1981) ideal test. Authenticity, interactiveness, and impact are also considered qualities that add to the overall usefulness of the test (Brown & Abeywickrama, 2010; Douglas, 2000). Although the degree of each of these qualities may differ to fit the purpose, they cannot be evaluated independently. Nevertheless, the aim of the current study was to maximize the application of each quality in our test.

Test Reliability

Several aspects of reliability were considered to accommodate the wide range of language elements in the course syllabus, and the researcher increased the number of task types and questions compared to the previous midterm test. For the scoring to be objective, mostly multiple-choice testing was used. When the external factors related to the test papers are considered, utmost attention is paid to eliminating any kind of ambiguous wording of instructions. The exam setting was a familiar, quiet, well-lit classroom in the students' own department building. Following the scoring of the written component of the exam, two teachers then checked for the inter-rater reliability put forward by Miles and Huberman's formula (1994). The result was 85%, which is considered acceptable. Thereafter, the Kuder-Richardson formula (KR20) was used to determine internal consistency, and a satisfactory result of 0.697 was obtained.

Test Validity

The textbook used in the Basic English course is prepared for university students and young adults, following CEFR principles. It adopts a multi-strand syllabus and a skills-based approach built on communicative language teaching. Construct validity is achieved by defining constructs consistent with the purpose of the test and designing test tasks that will facilitate the researchers making inferences about language ability that match these definitions. What is covered in the classroom is reflected in the test. A multiple-choice format is preferred as the dominant test type for practical reasons, as it lends itself to covering many objectives. Grammar and vocabulary items were contextualized. Each module in the course book was designed to integrate all four skills. However, listening and speaking are not tested due to insufficient facilities and equipment. Lack of time and staff negates any testing for speaking, either. For content validity, a table of specifications describing the language areas or skills with their percentage weights was identified as the first step. Concerning face validity, to refrain from a negative impression that will affect student motivation, long and unclear instructions were eliminated. Two teachers and two research assistants were asked for their expert comments on their impression of the test.

Test Authenticity

The course book followed is relevant to students' needs in that it addresses the appropriate age group. The topics are carefully selected and are all relevant to the lives of university students. The characters, situations, and events in the questions are all connected to life on campus. This contributes a great deal to the authenticity of the test in that there is a correspondence between the characteristics of the target language's use context and the activities in the book. Authenticity has been defined in several ways in language testing circles. Some take it as face validity, namely, the test's appealing power on the test takers. However, this view is critical, as Davies (1997) maintains that a test that appeals to test writers may be different from one that appeals to students and teachers. Another common and simple definition of authenticity is 'real-life-like' language use (Hoekje & Linnell, 1996; Lewkowicz, 2000; Pill, 2016; Wu & Stansfield, 2001). The researcher utilized CEFR skill descriptions as a route map. In the initial part of the exam, the same discourse related to the daily routines of people working in different occupations was used as that featured in the textbook. This part was prepared to activate the content of what was spoken or written frequently in real life and during the course. Hence, utmost attention was paid to contextualizing multiple-choice questions. Consequently, the test items have a moderate level of authenticity.

Test Interactiveness

The test can be regarded as relatively interactive. The students' knowledge, metacognitive strategies, topical knowledge, and affective states determine the degree of interactiveness of a language test. In the first part of the exam, an excerpt about a person's everyday life was provided with blanks to test grammar and vocabulary. The purpose of constructing those items was to require the students to use language with reference to the world in which they live and activate their schemata and topical knowledge. In the situational dialogues section, the aim was to assess students' degree of mastery in functional language. Interactiveness was relatively high since the test task reflected the communicative use of the language to express and interpret meaning in terms of test takers' experience of the real world.

Test Impact

The stakeholders in the test were the students, researchers, language instructors, and departmental academic staff delivering the course. The scores were the students' final grades, which had a 50% effect on their total grade. This was the final and largest component of the total evaluation, and they would either pass or fail the course based on their scores. Considering the impact on instruction, the results of the test indicated that supplementary reading materials appropriate to their level, such as a compiled reading pack, are necessary. Besides, the scores of the test served diagnostic purposes. Especially in some groups, the scores revealed that particular areas in the syllabus needed remedial activities regarding mother tongue interference in some structural points.

Test Practicality

For this achievement test, practicality was one of the most significant aspects of its usefulness, as tests of this type are fairly demanding in terms of cost. The test in this study was cost-effective in that it required only the exam papers to be printed and no other preparation. A legible standard font size and style were chosen. Multiple choice was chosen for ease of scoring in most parts of the test. Concerning practicality, students were asked to circle the correct alternative on the question booklet and not on a separate sheet, which made it easy to answer. Nevertheless, the researcher needed to pay extra attention when scoring the questions one by one. In the administration stage, research assistants from the departments whose students took the exam were the invigilators. When it comes to practicality concerning the content of the test, the high cost of using communicative items in all parts of the test inevitably limited their use. Thus, the reason for using a mostly objective type of testing and objective question types, such as multiple-choice questions, is to avoid the scorer having to spend too much time on each paper. Add results and findings here. Add results and findings here. Add results and findings here. Add results and findings here. Add results and findings here. Add results and findings here. Add results and findings here. Add results and findings here. Add results and findings here. Add results and findings here.

Discussion and Conclusions

The current study aimed to develop a summative test for a compulsory Basic English course delivered to all freshmen students attending Turkish-medium universities. Various stages of test construction include writing test specifications, writing test items, moderating test items, standardizing the scoring key, administering the test, monitoring quality (Item Analysis), and refining and finalizing the test. The test was prepared to reflect classroom instruction, and almost all the areas covered in the course were included in the test except speaking and listening activities. Although these were integral to the course, they were not included in the exam due to time and cost efficiency concerns. However, to approximate classroom instruction and the test as much as possible, dialogue questions were added, which may be considered an indirect way of testing speaking. The utmost care was given to the wording of the test questions. The test was used for grading purposes and provided feedback on the achievement of the students. The test analysis process also served diagnostic purposes. An action plan was designed for the areas that needed remedial work. One of the aims of the study was to initiate action to form a question bank for this course. Course instructors were invited to exchange ideas on this issue. In general, there was a positive attitude towards forming a question bank by continuing the example in this context.

The present study depicts the preparation process of a summative test and the analysis of the results in detail. The problems identified can be symptomatic of deeper reasons underlying them; hence, this study bears significance in that the requirement for 'good' test design and administration is usually based on theoretical and normative literature; however, the exploration of an actual test preparation, administration, and analysis practice at institutions similar to the one in this study, such as Turkish-medium state universities in Turkiye, may shed light on good practice for the benefit of all practitioners teaching and assessing freshman English courses, which are currently a required course at all higher-level educational institutions in the country.

The researcher reached several conclusions. Firstly, the test can be regarded as a good test characterized by the test's usefulness features, which were mapped according to Bachman and Palmer's (1996) Test

Development Model. Secondly, it provides a tangible resource that instructors can use to enhance their teaching methods and improve their content. Throughout the development of the test, the researcher collaborated with teachers and gained insights from them regarding their increased awareness of the significance of analysis in test quality. They stated their opinions on how collaboration for test preparation and following a set of guidelines for a unified model as suggested in the study could be put into practice and the needs of all departments taken into consideration. Furthermore, they stated that the test set an example to shape their classroom instruction and testing, complementing each other.

Tests may have positive feedback, as the teachers stated, on the content of classroom instruction. Some doubts have been raised as to whether teachers' methodology is affected by teacher-produced tests. Studies regarding the test quality of teacher-produced tests coincide with the researchers' claim when they initiated work on all the stages of the test; that is, no single test can meet all the needs, and no conclusion can be drawn for classroom instruction methodology. However, the researcher's experience throughout the study is in line with the studies that are for teacher-produced tests (Ahmad & Rao, 2012; Coniam, 2009; Galluzzo, 2005; Lan & Fan, 2019; Marzano, 2020; Mertler, 2004; Shohamy, 2020), which report that merely an exam on its own cannot reinforce an approach to teaching until the efforts of material designers, testers, and teachers are united by shared values. Therefore, the study is significant in that it brought the immediate stakeholders together to work on the development and analysis of the test's usefulness based on a model that utilized the teaching materials used on the course. If the researcher did not have a fundamental understanding of the nature of communication between teachers and their students, creating detailed test specifications, achieving a fit between those specifications and test items, and analyzing the test's usefulness would not have brought about a desire amongst stakeholders to apply the same model to create future tests for the course test bank.

Limitations

Data collection (test development) rests on questions generated through a one-shot achievement test. During the stages of preparation, administration, and grading, test specifications, items, and their analysis data were obtained from the participants in the study. Different tests distributed at several intervals could be used to increase the reliability of the data; however, formal constraints from the faculty, namely, the official allocation of tests in allocated settings for the mid-term and final exam periods. Data gathering can be expanded by following the same stages in a longitudinal study. The research is only limited to the so-called departments given above at a state university. More and varied data can be gathered from different departments, and they would have implications for further quantitative and qualitative studies. To have a deeper understanding and a mutual agreement with the teachers delivering the courses, personal conversations were conducted in office settings. Adopting mixed-methods approaches would yield more profound results.

Acknowledgements

The researcher acknowledges that the test development and mapping it according to a model would not be possible without the help of teachers delivering the course, faculty invigilators and students' participation.

Conflicts of Interest

The author declares that she has no personal and financial conflict of interest associated with this publication to disclose. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Ethical Approval

Ethical permission (2021-232) was obtained from Necmettin Erbakan University Ethical Commission for this research.

References

Abraham, R. G., Chapelle, C. A. (1992). The meaning of cloze test scores: An item difficulty perspective. Modern Language Journal, 76(4), 468–479. https://doi.org/10.1111/j.1540-4781.1992.tb05394.x

Adams, R. J., Griffin, P. E., Martin, L. (1987). A latent trait method for measuring a dimension in second language proficiency. *Language Testing. 4*(1), 9–28.

Ahmad, S., Rao. (2012). A Review of the Pedagogical Implications of Examination Washback. *Research on Humanities and Social Sciences, 2(*7), 11- 20.

Alan, B. (2003). Novice teachers' perceptions of an in-service teacher training course at Anadolu University. Unpublished master's thesis. Bilkent University, Ankara.

Alderson, J. C., Hamp-Lyons, L. (1996). TOEFL preparation courses: a study of washback. *Language Testing. 13*(3), 280–97.

Aschbacher, P. R. (1991). Performance assessment: state of activity, interest, and concerns. *Applied Measurement in Education, 4*, 275–288.

Atay, D. (2008). Teacher research for professional development. *ELT Journal, 62*(2), 139-147.

Bachman, L. F. (1991). 'What Does Language Testing Have to Offer?' *TESOL Quarterly, 25*, 671-704.

Bachman, L. F., Lynch, B. K. & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing 12*, 238-257.

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

Bachman, L.F. & Eignor, D.R. (1997). Recent advances in quantitative test analysis. In Clapham, C. and Corson, D., editors, Encyclopedia of language and education. Volume 7: Language testing and assessment. Kluwer Academic Publishers, 227–242.

Bachman, L.F. (1997). Generalizability theory. In Clapham, C. and Corson, D., editors, Encyclopedia of language and education. Volume 7: Language testing and assessment. Kluwer Academic Publishers, 255–262.

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

Bailey, K.M. (1996). Working for washback: a review of the washback concept in language testing. *Language Testing. 13*(3), 257–79.

Balbay, S., & Pamuk, I., Temir, T. Doğan, C. (2018). Issues in pre-service and in-service teacher training programs for university English instructors in Turkey. *Journal of Language and Linguistic Studies, 14*(2), 48-60.

Ballıdağ, S. (2020). Exploring the Language Assessment Literacy of Turkish In-service EFL Teachers. https://doi.org/10.06.2020.

Ballıdağ, S., & Inan 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

Banerjee, J., Luoma, S. (1997). Qualitative approaches to test validation. In Clapham, C. and Corson, D., editors, Encyclopedia of language and education. Volume 7: Language testing and assessment. Dordrecht: Kluwer Academic, 275–287.

Bell, R. T. (1981). An introduction to applied linguistics. Batsford Academic Ltd.

Bolt, R.F. (1992). Cross Validation of item response curve models using TOEFL data. *Language Testing. 9*, 79–95.

Bonner, S. M., Torres Rivera, C., & Chen, P. P. (2018). Standards and assessment: Coherence from the teacher's perspective. *Educational Assessment, Evaluation and Accountability, 30*,71-92.https://doi.org/10.1007/s11092-017-9272-2.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing. 12*(1), 1–15.

Brown, H. D. (2004). Language assessment: Principles and classroom practices. Longman.

Brown, J. D. (2013). My twenty-five years of cloze testing research: So, what? *International Journal of Language Studies, 7*(1), 1-32.

Brown, H. D., Abeywickrama, P. (2010). Language Assessment. Principles and Classroom Practices. White Plains.

Bull, M. and Yoneda, M. (2012). Designing assessment tools: The principles of language assessment. *Humanities and Social Sciences, 60,* 41－49.

Canale, M. & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 8*, 67–84.

Cheng, L. (1999). Changing assessment: washback on teacher perceptions and actions. *Teaching and Teacher Education. 15*, 253–271.

Cizek, G. J. (2000). Pockets of resistance in the assessment revolution. *Educational Measurement: Issues and Practice, 19*(2), 16-23. https://doi.org/10.1111/J.1745-3992.2000.TB00026.X

Clapham, C. (1996). The development of IELTS: a study of the effect of background knowledge on reading comprehension. University of Cambridge Local Examinations Syndicate/Cambridge University Press.

Cochran, J. L., McCallum, R. S., Bell, S. M. (2010). Three A's: How Do Attributions, Attitudes, and Aptitude Contribute to Foreign Language Learning? *Foreign Language Annals, 43*(4), 566–582.

Coniam, D. (2009). Investigating the quality of teacher-produced tests for EFL students and the effects of training in test development principles and practices on improving test quality. *System,37,* 226–242.

Coombs, A., DeLuca, C., LaPointe-McEwan, D., & Chalas, A. (2018). Changing approaches to classroom assessment: An empirical study across teacher career stages. *Teaching and Teacher Education,71,* 134-144. https://doi.org/10.1016/J.TATE.2017.12.010.

Darwesh, A. J. A. (2010). Cloze tests: An integrative approach. *Journal of the College of Basic Education, 15*(64), 105-116.

Davies, A. (1997). The construction of language tests. Oxford University Press.

Douglas, D. (2000). Assessing language for specific purposes: theory and practice. Cambridge University Press.

Ekşi, G. (2010). An assessment of the professional development needs of English language instructors working at a state university. (Unpublished master's thesis). Middle East Technical University, Ankara.

Fanrong, W., & Bin, S. (2022). Language Assessment Literacy of Teachers. In Frontiers in Psychology (Vol. 13). Frontiers Media S.A. https://doi.org/10.3389/fpsyg.2022.864582.

Fulcher, G. (2012). Assessment literacy for the language classroom, *Language Assessment Quarterly, 9(*2), 113-132, https://doi.org/10.1080/15434303.2011.642041.

Galluzzo, G. R. (2005). Performance assessment and renewing teacher education the possibilities of the NBPTS standards. The Clearing House: *A Journal of EducationalStrategies, Issues,and Ideas, 78*(4), 142-145. https://doi.org/10.3200/TCHS.78.4.142-145.

Ginther, A., Stevens, J. (1998). Language background and ethnicity, and the internal construct validity of the Advanced Placement Spanish Language Examination. In Kunnan, A.J., editor, Validation in language assessment. Lawrence Erlbaum, 169–94.

Giraldo, F., & Murcia Quintero, D. (2019). Language Assessment Literacy and the Professional Development of Pre-Service Language Teachers. *Colombian Applied Linguistics Journal, 21*(2), 243–259. https://doi.org/10.14483/22487085.14514.

Grondlund, N. E., Linn, R. L. (1990). Measurement and Evaluation in Teaching. Macmillan.

Gruba, P. and Corbel, C. (1997). Computer-based testing. In Clapham, C., Corson, D., editors, Language testing and assessment. Language testing and assessment. Dordrecht: Kluwer Academic, 141–149.

Gültekin, İ. (2007). The analysis of the perceptions of English language instructors at TOBB University of Economics and Technology regarding inset content. Unpublished master's thesis. Middle East Technical University, Ankara.

Halliday, M. A. K. (1973). Explorations in the functions of language. Edward Arnold.

Heaton, J. B. (1990). Writing English language tests. Longman.

Hoekje, B. & Linnell, K. (1996). Authenticity in language testing: Evaluating language tests for international teaching assistants. *TESOL Quarterly, 28*(1), 103-126. https://doi.org/10.2307/3587201.

Horwitz, E. (2001). Language anxiety and achievement. *Annual Review of Applied Linguistics. 21*(1), 112-126.

Höl, D. (2023). Standardized testing in Turkey: EFL teachers' perceptions and experiences on Cambridge Young Learner Exams (YLE). International Journal of Curriculum and Instruction. 15(2), 984-1007.

Hughes, A. (2010). Testing for language teachers. CUP.

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research. 60*, 237–63.

Hymes, D. H. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), Sociolinguistics. Penguin Books.

Kirschner, M., Wexler, C., Specto, E. (1992). Avoiding obstacles to student comprehension of test questions. *TESOL Quarterly, 26,* 537-556.

Kirschner, M., Wexler, C., Specto, E. (1996). A Teacher Education Workshop on the Construction of EFL Tests and Materials. *TESOL Quarterly, 30*, 85-111.

Kirkgöz, Y. (2008). Globalization and English Language Policy in Turkey. *Educational Policy, 23* (5), 663-684. https://doi: 10.1177/0895904808316319.

Köksal, D., Erten, İ. H., Zehir Topkaya, E., Yavuz, A., Yüksel, G., Aksu, İ. E. A., Şirin, E. (2006). English Course for Young Adults: Campus Life, Nobel Yayın Dağıtım.

Krashen, S. D. (1982). Principles and practice in second language acquisition. Pergamon Press.

Lado, R. (1964). Language testing: The construction and use of foreign language tests. McGraw-Hill.

Lan, C., & Fan, S. (2019). Developing classroom-based language assessment literacy for in-service EFL teachers: The gaps. *Studies in Educational Evaluation, 61*, 112–122. https://doi.org/10.1016/j.stueduc.2019.03.003

Lantolf, J., Frawley, W. (1985). Oral proficiency testing: a critical analysis. *Modern Language Journal 69*, 337–45.

Latif, M. W., & Wasim, A. (2022). Teacher beliefs, personal theories and conceptions of assessment literacy—a tertiary EFL perspective. *Language Testing in Asia, 12*(1). https://doi.org/10.1186/s40468-022-00158-5

Lewkowicz, J. (2000). Authenticity in language testing: Some outstanding questions. *Language Testing. 17*(1), 43-64.

Llosa, L. (2011). Standards-based classroom assessments of English proficiency: A review of issues, current developments, and future directions for research. *Language Testing. 28* (3), 367-382.

Lumley, T., McNamara, T. F. (1995). Rater characteristics and rater bias: implications for training. *Language Testing, 12*(1), 54–71.

Luoma, S. (2001). Test review. *Language Testing, 18*(2) 225–23. https://doi.org/10.1177/026553220101800207
Lynch, B. K., Davidson, F., Henning, G. (1988). Person dimensionality in language test validation. *Language Testing, 5*(2), 206–19.
Madsen, H. (1983). Techniques in testing. Oxford Pub.
Marzano, R. J. (2000). Transforming classroom grading. Association for Supervision and Curriculum Development.
McNamara, T.F. (1991). Test dimensionality: IRT analysis of an ESP listening test. *Language Testing, 8*(2), 139–59.
Mertler, C. A. (2004). Secondary teachers'assessment literacy: Does classroom experience make a difference?American Secondary Education,33(1),49-64.https://www.jstor.org/stable/41064623
Miles, M. B. and Huberman, M. A. (1994). An expanded sourcebook: Qualitative data analysis. Sage.
O'Loughlin, K. (2013). Developing the assessment literacy of university proficiency test users. *Language Testing, 30*(3), 363-380. https://doi.org/10.1177/0265532213480336.
Oller, J. W., Jr. (1976). Evidence for a general language proficiency factor. *DieNeuren Sprachen, 76*, 165–174.
Ölmezer-Öztürk, E., & Aydin, B. (2019). Investigating language assessment knowledge of efl teachers. *Hacettepe Egitim Dergisi, 34*(3), 602–620. https://doi.org/10.16986/HUJE.2018043465.
Pill, J. (2016). Drawing on indigenous criteria for more authentic assessment in a specific-purpose language test: Health professionals interacting with patients. *Language Testing, 33*(2), 175–193.
Pollitt, A. (1997). Rasch measurement in latent trait models. In Clapham, C. and Corson, D., editors, Encyclopedia of language and education. Volume 7: Language testing and assessment. Kluwer Academic, 243–54.
Read, J., Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing, 18*(1), 1-32.
Purpura, J. E. (1997). An analysis of the relationships between test takers' cognitive and metacognitive strategy use and second language test performance. *Language Learning, 47*, 289–325.
Richards, J. (2002). 30 Years of TEFL/TESL experience: A personal reflection. *RELC Journal, 33*, 1-35.
Sahlberg, P. (2006). Education reform for raising economic competitiveness. *Journal of Educational Change,7(*4), 259-287. https://doi.org/10.1007/S10833-005-4884-6.
Sariyildiz, G. (2018). Department of Foreign Language Education English Language Teaching A Study Into Language Assessment Literacy Of Preservice English As A Foreign Language Teachers In Turkish Context.
Sasaki, M. (1996). Second language proficiency, foreign language aptitude, and intelligence: quantitative and qualitative analyses. Peter Lang.
Shohamy, E. (2001). The power of tests. Longman.
Shohamy, E. (2020). The Power of Tests: A critical perspective on the uses of language tests (1st ed.). Routledge.
Shohamy, E., Donitsa-Schmidt, S., Ferman, I. (1997). Test impact revisited: washback effect over time. *Language Testing 13*(3), 298– 317.
Spolsky, B. (1978). Introduction: Linguists and language testers. In Spolsky, B. (ed.) Advances in language testing research: Approaches to language testing. Vol. 2. Washington, DC: Center for Applied Linguistics.
Spolsky, B. (2002). Prospects for the survival of the Navajo language. *Anthropology and Education Quarterly, 33*(2), 139-162.
Stansfield CW. Lecture (2008). Where we have been and where we should go. *Language Testing, 25*(3), 311-326. https://doi:10.1177/0265532208090155.
Şentuna, E. (2002). The interests of EFL instructors in Turkey regarding inset content. Unpublished master's thesis. Bilkent University, Ankara.
Terwilliger, J. (1998). Semantics, psychometrics, and assessment reform: a close look at 'authentic' assessments. *Educational Researcher, 26*(8), 24–27.
Tomak, B., Karaman, A. C. (2013). Mentoring in a professional development program for novice teachers at a state university in Turkey: a qualitative inquiry. *The International Journal of Research in Teacher Education, 4* (2), 1-13.
Ur, P. (1996). A course in language teaching: Practice and theory. CUP.
Vogt, K., Tsagari, D., & Spanoudis, G. (2020). What Do Teachers Think They Want? A Comparative Study of In-Service Language Teachers' Beliefs on LAL Training Needs. *Language Assessment Quarterly*, 386–409. https://doi.org/10.1080/15434303.2020.1781128.
Wall, D. (1996). Introducing new tests into traditional systems: insights from general education and from innovation theory. *Language Testing, 13*(3), 334–54.
Wall, D., Alderson, J. C. (1993). Examining washback: the Sri Lankan impact study. *Language Testing, 10*, 41–69.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263–67.

Willis, D. (2003). Rules, patterns, and words. Cambridge University Press.

Wu, W. M., Stansfield, C. W. (2001). Towards authenticity of task in test development. Language Testing, 18(2), 187-206. American Psychological Association. (2010). Publication manual of the American Psychological Association (6th ed.). Washington, DC: American Psychological Association.

Yamtim, V., & Wongwanich, S. (2014). A Study of Classroom Assessment Literacy of Primary School Teachers. *Procedia Social and Behavioral Sciences, 116,* 2998–3004. https://doi.org/10.1016/j.sbspro.2014.01.696

Appendix

Bachman and Palmer's Test Development Stages (1996, p. 87)