Sakarya University Journal of Education

ISSN : 2146-7455
Publisher : Sakarya University

Vol. 14, No. 3, 473-495, 2024
DOI: https://doi.org/10.19126/suje.1447028

Research Article

Comparison of Item Difficulty Analyses of Exams Used in Teaching Turkish as a Foreign Language with Instructors' Perceptions of Item Difficulty

Funda Keskin[1]*
Seçil Alaca[2]

[1] Sakarya University, Institute of Educational Sciences, Sakarya, Türkiye, funda.keskin@istanbul.edu.tr

[2] Istanbul University, Language Center, İstanbul, Türkiye, secilalaca@gmail.com

*Corresponding author

Received: 04.03.2024
Accepted: 01.10.2024
Available Online: 04.11.2024

Abstract: In numerous studies focusing on assessment and evaluation of teaching Turkish as a foreign language, researchers have frequently identified issues related to the standardization and low validity and reliability of exams. Addressing these issues and investigating the underlying causes is paramount. Given the development of assessment tools by Turkish language teaching centers are typically the responsibility of instructors, it is essential to understand their perspectives regarding these tools. This study aimed to evaluate the perceptions of instructors concerning item difficulty in the context of teaching Turkish as a foreign language. Initially, item analyses were conducted on reading tests included in assessment tools designed by a Turkish language teaching center for B1, B2, and C1 proficiency levels. Instructors from various Turkish language teaching centers were asked to evaluate item difficulty through a prepared questionnaire. Data regarding instructors educational backgrounds, experiences, and involvement in exam creation were collected. Various analytical methods were employed to examine and interpret the obtained data. Item analysis results of examined tests were compared with instructors' perceptions of difficulty using fit analysis. Accuracy of instructors' item difficulty estimates was calculated for each instructor using Error Matrix, and success rates determined. To identify the effects of instructors' characteristics on item difficulty estimation, t-test and ANOVA analyses were performed. These analyses results were interpreted alongside item analyses, and recommendations provided to enhance the assessment and evaluation literacy of instructors teaching Turkish as a foreign language.

Keywords: Turkish as a Foreign Language, Item Analysis, Item Difficulty Perception, Assessment in Teaching Turkish as a Foreign Language

1. Introduction

In language tests, item writers can be an extremely important aspect which directly affects test validity, such as their impact on test and test specification development. No matter how high the theoretical validity of a developed test, item writers competence can directly affect test validity. Therefore, the performance of item writers and their accurate guidance is an important stage of test development. To improve item writer performance, it is necessary to first identify the situation. Importantly, the appropriate guidance needed by question writers is enhanced by the performance of item writers. One of the best ways to determine item writer performance is through actual item writing and piloting, even though this is not always practical. For this reason, several studies have relied on expert judgement of item difficulty to determine performance indicators (Fergadiotis et all., 2019; Hambleton & Jirka, 2006; Sydorenko, 2011; Wise et all., 2009).

In the test development process, having knowledge regarding item difficulty is crucial. For a particular group of test participants, the difficulty of items can be determined fairly accurately following pilot testing. However, it is necessary to understand what makes items more or less difficult while being developed and prior to being piloted. A common practice in test development is to provide item writers with item-level descriptors, which are usually grounded in previous research related to predictors of item difficulty. For example, previous research has identified various factors that affect item difficulty including negation in the stems (e.g., Hambleton & Jirka, 2006), topic familiarity (e.g., Freedle & Kostin, 1999), or lexical knowledge (e.g., Rupp, Garcia, & Jamieson, 2001). One of the purposes of such research is to increase item-writing process efficiency by providing specific guidelines and item-level descriptors to item writers (Kostin, 2004). Additionally, further studies have expanded on these factors. For

Cite as (APA 7): Keskin, F., & Alaca, S. (2024). Comparison of item difficulty analyses of exams used in teaching Turkish as a foreign language with instructors' perceptions of item difficulty. *Sakarya University Journal of Education,* 14(3), 473-495. https://doi.org/10.19126/suje.1447028

This is an open access paper distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International License.

instance, Bachman (2002) highlighted even experienced item writers often struggle with accurately predicting item difficulty, emphasizing the need for extensive training and detailed guidelines. Similarly, Alderson (1993) found judges are better at predicting the difficulty of reading comprehension items compared to other types, such as cloze tests, indicating variability in difficulty prediction across different item formats. Moreover, Bejar (1983) demonstrated the accuracy of item difficulty estimates can be significantly improved through use of anchor-based methods and training, though these improvements are not always sufficient to entirely replace empirical pretesting. Furthermore, Shohamy (1984) indicated the type of response format (i.e., multiple-choice vs. true-false) significantly impacts item difficulty, with multiple-choice items generally being more challenging due to lower probability of guessing correctly. This body of research collectively aimed to refine the item development process by providing item writers with robust, empirically validated tools, and guidelines which enhance the predictive accuracy of item difficulty. By integrating these findings, test developers can create more reliable and valid assessments to more accurately measure intended skills and knowledge.

Alderson, Clapman, and Wall (1995) argued that no matter how well a test is designed, it can be quite challenging to determine whether the items are appropriate without first piloting them on learners. Even experienced teachers and test experts often disagree on what a specific item measures or how difficult it may be. Therefore, piloting is essential to assess test validity and reliability. Boylu (2019) notes that while item analysis might seem challenging, instructors can easily conduct difficulty and discrimination analyses. However, many educators feel they lack proficiency in areas such as "calculating item discrimination", "determining which items to include based on discrimination index scores", and "calculating test reliability and validity" (Altıntaş, 2022; Boylu, 2019).

Exams created by Turkish Language Teaching Centres (TLTC) often lack validity and reliability due to instructors' limited knowledge regarding item writing and language assessment (Gedik, 2017; Işıkoğlu, 2015). This leads to inaccurate assessments which can negatively affect students (Kutlu et al., 2010). Although the inadequacy of assessment tools and instructors' difficulties in applying language assessment principles have been identified, studies focusing on the root cause of these issues remain insufficient. For example, current research measures instructors' theoretical knowledge through surveys and tests (Boylu, 2019; Çavuşoğlu & Işık, 2021; Karagöl, 2020; Mercan & Göktaş, 2023; Özdemir, 2023; Sertdemir, 2021), but independent studies are needed to address specific stages of assessment tool development. Furthermore, short-term targeted training can offer long-term benefits by addressing specific problem areas.

2. Study Purpose

In language exams, it is important for item writers to know the principles of assessment, write questions considering content validity, and comprehend the assessment objectives of items. An item writer should also know whether an item is appropriate for the target group. In this case, item writers should to be able to determine the difficulty levels of items they produce as well as content validity. It is recommended in the literature to assemble a test with items of varying difficulty which address all ability levels within a target population. In other words, in a test designed to assess reading skills in proficiency and placement exams, the item difficulties of all items should not cluster around a certain difficulty level. For example, a test consisting of only easy or difficult items could jeopardize the validity, reliability, and discriminatory power of the test. Therefore, it is important the item difficulty perceptions of item writers align with the results of item analyses.

Therefore, the aim of this study included, first, to compare the item difficulty perceptions of instructors and item writers in the field of teaching Turkish as a foreign language with the item difficulty indices obtained from item analyses, and second, to determine whether their predictions regarding the item difficulties were correct or not.

Additionally, another aim of this study was the analysis of different factors which influence the consistency of teachers' estimates regarding item difficulty.

If it was determined the item difficulty predictions of instructors were not consistent with the item analyses, therefore, an additional aim of the study, which was dependent on the data results, was to make suggestions on how item writers can improve in this regard.

3. Method

In the study, descriptive survey method based on quantitative variables was used. The descriptive survey method aims to describe an existing situation regarding a specific group through use of a questionnaire. In this survey method, which is frequently used in educational research, data are collected from a specific group at a specific time. Descriptive survey method provides information about the behaviors, ideas, beliefs, knowledge, and so forth of the participants in the survey. Through the collected data – descriptions, comparisons, and classifications – can be made about the questions related to the research (Cohen et al., 2000).

Based on the descriptive survey method, our study aim was to determine the surveyed instructors perceptions of item difficulty. In this context, first, item analyses of tests applied to various groups at different language levels were conducted. Then, a survey was administered to a group of 51 instructors with varying years of experience, levels of education, and education program, asking them to estimate the difficulty level of test items. According to our survey results, instructors' difficulty estimates of the items and difficulty levels of the items according to our analyses were compared, and it was determined to what extent instructors estimates regarding item difficulty levels were compatible.

Considering the literature highlights that the predictions of item writers are more compatible, especially in reading skills and multiple-choice items, we first wanted to determine overall consistency. For this reason, agreement analysis was conducted to provide a more general view to analyze the consistency between item difficulties and instructors' predictions. First, we looked at prediction agreement according to the level for the Common European Framework of Reference for Languages (CEFR). As a result of the low agreement rates in the results obtained from the agreement analysis, we decided to expand the analysis. In addition to agreement analyses, to analyze inter-rater reliability among instructors' predictions, a Fleiss' Kappa analysis was performed (Nichols et al., 2010). As the agreement in this analysis was also low, to understand common patterns and types of errors made by all instructors, we decided to apply confusion matrix for each instructor. Even though confusion matrix is a performance measurement tool for machine learning classification models, the fact that our data was quite suitable for this model persuaded us to consider this analysis method could be useful for our evaluation.

Furthermore, to see whether the instructors' characteristics, which was another aim of our study, had an effect on their predictions, participants were grouped according to various characteristics and ANOVA analyses were performed using SPSS statistical analysis software.

3.1. Sample group

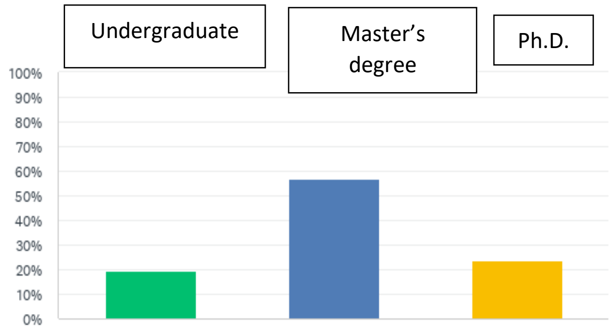
3.1.1. Demographic information

Instructors from various disciplines teach Turkish as a foreign language. The instructors working in TLTC's have also achieved different levels of education such as undergraduate, graduate, and doctorate degrees. In addition to these differences, there are a variety of certificate programs offered for teaching Turkish as a foreign language and some instructors participate in such programs. Thus, to observe whether such differences have an effect on prediction of item difficulty value, in the first stage of the questionnaire, the instructors were queried about their individual characteristics regarding the field of teaching Turkish as a foreign language.

Figure 1 and Figure 2 illustrate ratios related to participating instructors' educational status along with the university departments from which they graduated.

Figure 1

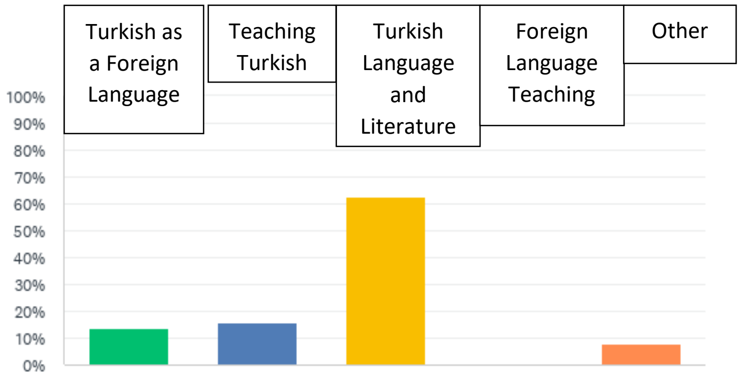
Instructors' Education Level



It can be seen in Figure 1 a majority of instructors (approximately 57%) have earned a Master's degree.

Figure 2

Education Programs from which Instructors Graduated

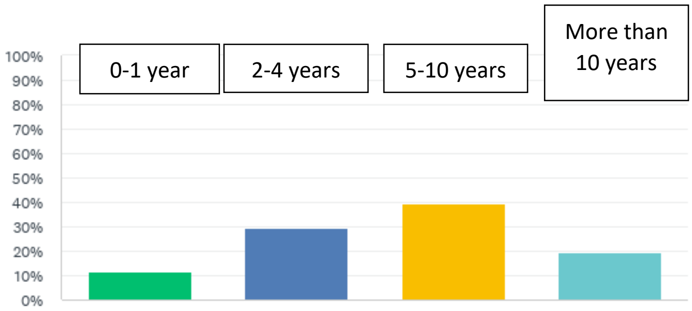


It can be seen in Figure 2 that 14% of participating instructors graduated from a teaching Turkish as a foreign language program, 16% from teaching Turkish, 64% from Turkish language and literature, and 8% from other departments such as Linguistics or elementary school teaching).

Instructors' experience in teaching Turkish as a foreign language was grouped according to 0-1, 2-4, 5-9, and 10 or more years.

Figure 3

Instructors Experience in Years



In Figure 3 it can be seen a majority of instructors (approximately 39%) have between two to five years of teaching experience. While, approximately 12% have zero to one year of teaching experience, 29% with two to four years teaching experience, and 20% with 10 or more years teaching experience.

In addition to demographic information presented in Figures 1, 2, and 3, 75% of instructors stated having received training specific to teaching Turkish as a foreign language. Along with this, 85% of instructors stated participating in certificate programs related to teaching Turkish as a foreign language, while 78% stated regularly preparing various exams.

4. Data Collection Methods

For the purpose of this study, two different data collection methodologies were used. The first was to analyze tests utilized in the TLTC. In the beginning stage of this study, the tests were analyzed, while in the second study stage, an item difficulty perception survey was developed and administered to instructors via Survey Monkey. Data obtained from the survey were used to access instructors' demographic information as well as to compare instructors' item difficulty perceptions with difficulty values from the item analysis. Next, the final study stage involved comparing correlations between instructors' item difficulty estimates and the item analyses from different analyses.

4.1. Item analysis

Within the scope of this study, first, item analyses of test sections assessing reading skills. prepared for different levels, were analyzed. For example, at the B1 level, two different tests were used: fill-in-the-blank, to measure grammatical accuracy, and a reading test to assess reading comprehension, which consisted of five multiple-choice (MC) items with each including three options. Similarly, at the B2 level, two tests were used: fill-in-the-blank test, prepared to measure grammatical accuracy, including the appropriate options, and a reading test aimed at assessing reading comprehension skills, which consisted of five multiple-choice (MC) items each with three options. Next, at the C1 level, only a fill-in-the-blank test was utilized to measure grammatical accuracy, vocabulary, and cohesion. Importantly, tests used at a TLTC as part of course achievement exams, was included as part of reading skills assessment. Item analysis for each test was carried out with data from exams administered in the TLTC. Information regarding the number of candidates and item types is presented in Table 1.

Table 1

Information Regarding Tests

Test	Candidate Numbers	Number of Items	Item Types
B1.1	81	10	MC / 3 options
B1.2	81	5	MC / 3 options
B2.1	86	10	MC / 3 options
B2.2	56	5	MC / 3 options
C1.1	59	10	MC / 3 options

Our test data were analyzed through the TAP program with item difficulty, item discrimination and biserial, point biserial correlations calculated for every test. Test reliability was based on the K20 values calculated for each test through TAP. Test reliability, mean item difficulty, and test discrimination rates are presented in Table 2.

Table 2

Information Regarding Reliability, Discrimination, and Mean Difficulty of Tests

Test	Test Reliability [K20(Alpha)]	Test Discrimination (Mean Point Biserial)	Mean Item Difficulty
B1.1	0.672	0.504	0.635
B1.2	0.555	0.600	0.548
B2.1	0.548	0.447	0.653
B2.2	0.585	0.613	0.550
C1	0.638	0.462	0.329

K20 Alpha values in Table 2 show test reliability. Tests with a K20 Alpha value higher than 0.6 have an acceptable reliability level. However, considering the small sample size, all tests in this study had only marginally acceptable reliability with values above or around .60.

Within the scope of this study, the theory proposed by Crocker and Algina (1986) was referred to as a basis for analyzing test items with items with a discrimination index greater than 0.40 determined as good, and items between 0.30 and 0.39 considered acceptable. However, items below 0.30 did not meet discrimination levels. Item difficulty was evaluated between 0 and 1 in the literature and items with 0.50 considered to have moderate difficulty. Importantly, it is generally recommended to include items of moderate difficulty in a test, yet to ensure discrimination, difficult and easy items should also be included in a test. Within the scope of the current study, items between 0.01 and 0.40 were considered difficult, items between 0.41 and 0.60 considered moderate difficulty, and items between 0.61 and 0.99 considered easy. Regardless, discrimination indices should also be considered when evaluating difficulty levels. In particular, items with difficulty values on the borderline such as 0.41 and 0.61 should be examined together with their discrimination indices and data interpreted, along with difficulty classification being made.

4.2. Data collection for item difficulty prediction

In the study’s second stage, a questionnaire was applied to instructors working in various institutions for teaching Turkish as a foreign language. In the questionnaire’s first stage, instructors were asked various questions such as the department from which they graduated, the degree they earned, amount of teaching experience, and whether they had prepared exams in the centers in which they worked. In the questionnaire’s second stage, instructors were asked about their item difficulty evaluations. Texts and items from the reading skills test were sent to the instructors and they were asked to determine item difficulty for each item as "difficult", "moderate", or "easy". The questionnaire prepared for instructors to complete was shared for them to access on "SurveyMonkey". Within the scope of this study, 51 instructors voluntarily completed the survey.

4.3. Analyzing consistency between difficulty level and difficulty prediction

The study’s final stage was to evaluate the relationship between instructors' item difficulty perceptions and item analyses. Since there were no studies identified in the literature regarding item difficulty predictions of instructors teaching Turkish as a foreign language, along with no data about the general consistency of instructors in this regard, it was determined to provide a preliminary overview. Therefore, to determine instructors’ agreement regarding item difficulties as well as the general view of whether a standard existed or not, we first examined instructors item difficulty predictions on the basis of inter-rater validity.

Consistency between item difficulty analyses and instructor evaluations was measured through method of agreement analysis. For example, the formula from Miles and Huberman (1994) was used in agreement analysis.

((Number of Agree / (Number of Agree + Number of Disagree))*100

The Miles and Huberman formula, generally used for inter-rater reliability, was utilized in this study to calculate agreement among instructor ratings. Instructors' item difficulty ratings were classified by taking into account the rating with the highest percentage. For example, an item rated by instructors as 36% (Low), 56% (Moderate), or 8% (High) was classified as "moderate". Furthermore, item difficulty levels were also coded by classifying them as difficult, moderate, or easy based on the item analysis results.

As a result of the low agreement and lack of consistency in the item difficulty predictions of instructors, a more detailed analysis was necessary. Thus, in addition to agreement analyses, as a means of analyzing inter-rater reliability among instructors' predictions, a Fleiss' Kappa analysis was performed (Nichols et al., 2010), and a Kappa score of 0.049 found. This score indicated only a slight agreement among instructors' predictions, indicating minimal consensus regarding item difficulty categorization. Additionally, instructors' predictions of item difficulty levels, including low (easy), moderate, and high (difficult) were converted into ordinal values; 1, 2, and, 3, respectively. Difficulty levels of items found through item analyses were also converted into ordinal values, where > .70 was coded as 1 (easy), .70-.30 as 2 (moderate), and < .30 as 3 (difficult).

As a result of low prediction success, it was decided to apply confusion matrix for each instructor. Even though confusion matrix is a performance measurement for machine learning classification models, the fact our data was quite suitable for this model convinced us to incorporate this analysis in the evaluation process. Thus, in this study, focused on instructor perceptions in language assessment, a different analysis methodology was conducted and in effect presented to the field. Indeed, the use of different analysis methods in the field of language assessment can likely increase opportunities for observation as well as to identify unforeseen problems and deficiencies. Based on the results of our use of confusion matrix, we first determined each instructor's item perception success along with interpreting the analysis results. Within the scope of this study, each instructor's matrix results were not shared, but one instructor's results are presented as an example as well as success percentage data of each instructor also presented in Appendix 1.

Next, to obtain detailed analysis data, ANOVA analyses were performed through SPSS. In ANOVA analyses, each instructor's item difficulty prediction (low, moderate, high) as well as item difficulties for the item analyses (easy, moderate, difficult) were coded. A paired sample t-test analysis was also conducted to determine whether prediction success varied at different CEFR levels. Demographic data obtained from the survey regarding instructors were also included in the analyses to understand whether different factors played a role in instructors' success in predicting item difficulty. Additionally, our analysis examined whether the success rates of trainers were related to their experience, level of education, department from which they graduated, having a foreign language teaching certificate, and/or preparing exams.

4.4 Ethical Principles

The ethics committee report for this study was obtained from the Istanbul University Rectorate Social and Human Sciences Research Ethics Committee with the decision dated 11.01.2024 and numbered 2024/19.

5. Findings and Discussion

Different analyses were carried out with differing data at different stages of this study. In the first stage, item analyses of reading tests for different language levels obtained from a TLTC were conducted. There were a different number of candidates and items at each level. In the second stage, the questionnaire percentages for instructors' item difficulty perceptions were determined as well as comparisons made regarding item difficulty values. In the third stage, agreement analysis of the instructors' difficulty ratings and item difficulty ratings were conducted.

5.1. Item analyses of B1, B2, and C1 tests

Following the item analyses, which was the first stage of this study, and conducted through TAP (Test Analysis Program), tables were prepared for analysis of data from each test. Then, these data were interpreted through the classification of theory and difficulty levels determined within the study's scope.

5.1.1. Item analyses of B1 level tests

Table 3

Item Analyses of B1.1 Test

B1.1 / 3 options MC /choosing the correct fill-in-the-blank answer							
Item	Correct Answers	Item difficulty	Discrimination index	High group	Low group	Biserial correlation	Point biserial correlation
Item 1	63	0.78	0.46	25 (0.93)	12 (0.46)	0.57	0.43
Item 2	55	0.68	0.58	25 (0.93)	9 (0.35)	0.56	0.40
Item 3	44	0.54	0.47	22 (0.81)	9 (0.35)	0.47	0.28
Item4	42	0.52	0.73	26 (0.96)	6 (0.23)	0.61	0.45
Item5	50	0.62	0.81	26 (0.96)	4 (0.15)	0.67	0.53
Item6	37	0.46	0.47	18 (0.67)	5 (0.19)	0.43	0.24
Item7	55	0.68	0.50	24 (0.89)	10 (0.38)	0.47	0.29
Item8	60	0.74	0.58	27 (1.00)	11 (0.42)	0.55	0.40
Item9*	54	0.67	0.28	21 (0.78)	13 (0.50)	0.38	0.19
Item10	54	0.67	0.39	25 (0.93)	14 (0.54)	0.35	0.15

In Table 3, it can be seen the discrimination index of Item 9 was below 0.30 and therefore its discrimination was not in the acceptable range. All items except Item 9 had good or acceptable discrimination indices. Thus, it can be seen from the analysis that Items 4 and 6 had moderate difficulty and all other items had low difficulty levels, that is, they should be considered as easy.

Table 4

Item Analyses of B1.2 Test

B1.2 / 3 options MC / reading comprehension							
Item	Correct Answers	Item difficulty	Discrimination index	High group	Low group	Biserial correlation	Point biserial correlation
Item 1	34	0.42	0.90	23 (0.96)	2 (0.06)	0.69	0.44
Item 2	53	0.65	0.54	22 (0.92)	12 (0.38)	0.57	0.28
Item 3	58	0.72	0.49	23 (0.96)	15(0.47)	0.59	0.33
Item 4	30	0.37	0.72	18 (0.75)	1 (0.03)	0.64	0.36
Item 5	47	0.58	0.56	21 (0.88)	10 (0.31)	0.51	0.19

In Table 4, it is seen the discrimination indices for all items in the test were valid. In terms of difficulty level, it was determined Item 4 was difficult, Item 1 was moderate (close to difficult), Item 5 was moderate, and Items 2 and 3 were easy.

5.1.2. Item analyses of B2 level tests

Table 5

Item Analyses of B2.1 Test

B2.1 / 3 options MC /choosing the correct fill-in-the-blank answer							
Item	Correct Answers	Item difficulty	Discrimination index	High group	Low group	Biserial correlation	Point biserial correlation
Item 1	65	0.76	0.32	30 (0.88)	14 (0.56)	0.36	0.15
Item 2	66	0.77	0.41	33 (0.97)	14 (0.56)	0.38	0.18
Item 3	65	0.76	0.42	32 (0.94)	13 (0.52)	0.47	0.27
Item 4	39	0.45	0.59	27 (0.79)	5 (0.20)	0.52	0.31
Item 5	72	0.84	0.34	32 (0.94)	15 (0.60)	0.48	0.32
Item 6	58	0.67	0.61	29 (0.85)	6 (0.24)	0.58	0.38
Item 7	38	0.44	0.43	20 (0.59)	4 (0.16)	0.39	0.15
Item 8	41	0.48	0.47	24 (0.71)	6 (0.24)	0.40	0.17
Item 9	68	0.79	0.46	32 (0.94)	12 (0.48)	0.47	0.29
Item 10	51	0.59	0.47	27 (0.79)	8 (0.32)	0.38	0.14

The discrimination indices for all items in Test 1 at the B2 level were above the acceptability value. In terms of difficulty level, Items 1, 2, 3, 5, 6, 9, and 10 were easy, while Items 4, 7, and 8 were of moderate difficulty. Although there were no items considered difficult within the test, Item 6 had a discrimination index of 0.61, and Item 4 had a discrimination index of 0.59 which stood out in terms of low group and high group correct answer rates.

Table 6

Item Analyses of B2.2 Test

B2.2 / 3 options MC / reading comprehension							
Item	Correct Answers	Item difficulty	Discrimination index	High group	Low group	Biserial correlation	Point biserial correlation
Item 1	37	0.66	0.59	20 (1.00)	11 (0.41)	0.63	0.38
Item 2	23	0.41	0.74	17 (0.85)	3 (0.11)	0.69	0.45
Item 3	31	0.55	0.73	19 (0.95)	6 (0.22)	0.66	0.41
Item 4	29	0.52	0.55	17 (0.85)	8 (0.30)	0.55	0.26
Item 5	34	0.61	0.43	16 (0.80)	10 (0.37)	0.52	0.23

All test items in Table 6 had acceptable discrimination indices. Although the difficulty level of Item 2 was considered in the intermediate group, it could be accepted as difficult due to its high discrimination index. Item 3 could be evaluated similarly and accepted as difficult. On the other hand, Item 4, which had a difficulty level of 0.52, was considered to be of moderate difficulty due to its discrimination index being 0.55.

5.1.3. Item analyses of C1 level tests

Table 7

Item Analyses of C1 Test

C1 / 3 options MC / choosing the correct fill-in-the-blank answer							
Item	Correct Answers	Item difficulty	Discrimination index	High group	Low group	Biserial correlation	Point biserial correlation
Item 1	56	0.95	0.12	28 (1.00)	15 (0.88)	0.42	0.29
Item 2	51	0.86	0.47	28 (1.00)	9 (0.53)	0.62	0.46
Item 3	53	0.90	0.24	28 (1.00)	13 (0.76)	0.36	0.17
Item 4	47	0.80	0.59	28 (1.00)	7 (0.41)	0.57	0.35
Item 5	50	0.85	0.41	28 (1.00)	10 (0.59)	0.59	0.40
Item 6*	58	0.98	0.00	28 (1.00)	17 (1.00)	-0.02	-0.10
Item 7	54	0.92	0.29	28 (1.00)	12 (0.71)	0.48	0.32
Item 8	51	0.86	0.47	28 (1.00)	9 (0.53)	0.69	0.54
Item 9	49	0.83	0.35	28 (1.00)	11 (0.65)	0.38	0.15
Item 10	51	0.86	0.35	28 (1.00)	11 (0.65)	0.53	0.34

In Table 7, it can be seen the difficulty level of all items at the C1 level were easy. Also, it was recognized the discrimination indices of Items 1, 3, and 7 were outside the acceptable level as well as Item 6 had no discrimination. In fact, it can be seen the item worked in reverse when the biserial correlation and point biserial correlation of Item 6 were examined. However, considering the fact 58 out of 59 candidates who participated in the test answered the item correctly, the assessment that the item worked in reverse should not be accepted as valid. As a matter of fact, only one candidate answered the item incorrectly. Considering the difficulty level of all items in the test was low and the discrimination indices were generally below the acceptable limit, it was not possible to distinguish between the upper and lower groups. In short, it can be interpreted the items in this test did not reach the measurement target. As a matter of fact, the use of a measurement tool in achievement tests, in which all items are easy, does not serve any measurement purpose. However, since the purpose of this study was to determine instructors' perception of item difficulty, there was a need to evaluate test items with low discrimination. For this reason, the C1 test was also included in the study and the instructors' evaluations of test items also examined.

Item analyses for B1, B2, and C1 level tests revealed variations in difficulty and discrimination indices. While the B1 and B2 tests performed within acceptable parameters, with most items being easy but displaying appropriate discrimination, the C1 level test showed significant inadequacies. Additionally, items in the C1 test were largely too easy, and several failed to meet discrimination standards, reflecting broader issues in language testing practices as identified by Özdemir and Eke (2023), Sertdemir (2021), and Şimşek (2016). These findings highlighted a need for improvement in item-writing practices, particularly for higher-level assessments, to ensure accurate measurement of student abilities, especially in areas requiring higher-order cognitive skills (Aydemir & Çiftçi, 2008; Oktay, 2015; Özcan & Akçan, 2010).

5.2. Findings related to instructors' perceptions of item difficulty

After the item analyses of the achievement tests analysed within the scope of this study was evaluated in the light of certain theories, an item difficulty rating questionnaire was prepared for instructors in the second study stage. In the survey, text and item levels were specified and instructors were asked to determine item difficulty as "low", "moderate", or "high" for each item. Data were collected through the "SurveyMonkey" website.

Instructors' item difficulty perceptions were analysed through the "SurveyMonkey" survey application. In the analysis, instructors' evaluations for each item were ranked according to all participating instructors. To determine agreement of instructors' perceptions with item difficulty levels, the percentages showing instructors' perceptions of difficulty levels for items in each test along with item difficulty values obtained from the item difficulty analysis were compared.

5.2.1. Comparison of instructors' perceptions of difficulty and item difficulty values for the reading test B1.1

Item analysis of the first reading test at the B1 level was conducted regarding the results of 81 candidates. The test consisted of 10 MC items with three options. Candidates were asked to choose the appropriate answers for the fill-in-the-blanks in a 450-word text. Difficulty values for the items and instructors' ratings are presented in Table 8.

Table 8

Instructors' Item Difficulty Perception Ratings and Item Difficulty Analyses for B1.1

	Rates of instructors' assessment of item difficult perception	Item difficulty
Item 1	34% Low 58% Moderate 12% High	0.78
Item2	32% Low 56% Moderate 12% High	0.68
Item3	22% Low 48% Moderate 30% High	0.54
Item4	30% Low 46% Moderate 24% High	0.52
Item5	12% Low 62% Moderate 26% High	0.62
Item6	26% Low 50% Moderate 24% High	0.46
Item7	46% Low 36% Moderate 18% High	0.68
Item8	26% Low 64% Moderate 10% High	0.74

Table 8(Continued)

Item9	44% Low 54% Moderate 2% High	0.67
Item10	26% Low 50% Moderate 24% High	0.67

5.2.2. Comparison of instructors' perceptions of difficulty and item difficulty values for reading test B1.2

Item analysis for the second reading test at the B1 level was conducted regarding the results of 81 candidates. In this test, there was a text consisting of 489 words and five MC items with three options for reading comprehension.

Table 9

Instructors' Item Difficulty Perception Rates and Item Difficulty Analyses for B1.2

	Rates of instructors' assessment of item difficulty perception	Item difficulty
Item 1	48% Low 36% M%oderate 16% High	0.42
Item 2	44% Low 44% Moderate 12% High	0.65
Item 3	62% Low 26% Moderate 12% High	0.72
Item 4	14% Low 52% Moderate 34% High	0.37
Item 5	54% Low 40% Moderate 6% High	0.58

Item 4 in the B1.2 test differed from other items with a difficulty level of 0.37. A majority of instructors rated Item 4 as the most difficult test item, at Moderate difficulty. Also, a majority of instructors considered Item 5, which had Moderate difficulty for analysis data, as an easy item.

5.2.3. Comparison of instructors' perceptions of difficulty and item difficulty values for reading test B2.1

The first reading test at the B2 level consisted of 10 fill-in-the-blanks in a text consisting of 56 words and 10 MC items with three options where the appropriate option was marked. Item analysis of the test was conducted with the data regarding 86 candidates.

Table 10

Instructors' Item Difficulty Perception Rates and Item Difficulty Analyses for B2.1

	Rates of instructors' assessment of item difficulty perception	Item difficulty
Item 1	36% Low 54% Moderate 10% High	0.76
Item 2	32% Low 56% Moderate 12% High	0.77
Item 3	40% Low 48% Moderate 12% High	0.76
Item 4	16% Low 48% Moderate 36% High	0.45
Item 5	36% Low 52% Moderate 12% High	0.84
Item 6	34% Low 56% Moderate 10% High	0.67
Item 7	10% Low 64% Moderate 26% High	0.44
Item 8	18% Low 46% Moderate 36% High	0.48
Item 9	16% Low 54% Moderate 30% High	0.79
Item 10	32% Low 50% Moderate 18% High	0.59

5.2.4. Comparison of instructors' perceptions of difficulty and item difficulty values for reading test B2.2

In the second reading test at the B2 level, data from 56 candidates were used for item analysis. In the test, there were five MC items with three options for reading comprehension regarding a text of 516 words.

Table 11

Instructors' Item Difficulty Perception Rates and Item Difficulty Analyses for B2.1

	Rates of instructors' assessment of item difficulty perception	Item difficulty
Item 1	48% Low 34% Moderate 18% High	0.66
Item 2	48% Low 38% Moderate 14% High	0.41
Item 3	34% Low 52% Moderate 14% High	0.55
Item 4	36% Low 46% Moderate 18% High	0.52
Item 5	48% Low 36% Moderate 16% High	0.61

Although Item 2 in the B2.1 test was the item with the highest difficulty level, it was largely evaluated as an Easy or Moderate difficulty item by instructors. A similar situation was also observed for Item 5.

5.2.5. Comparison of instructors' perceptions of difficulty and item difficulty values for reading test C1

Item analysis of one C1 level test provided usable data for our study. Item analysis of the second test shared by the TLTC could not be conducted due to all participants answering the items correctly. Therefore, item analysis could only be performed on the first test. Thus, data from 59 candidates were used for item analysis. The test included a reading text of 494 words and 10 MC items with three options with the instruction to fill in 10 blanks within the text for the right option.

Table 12

Instructors' Item Difficulty Perception Rates and Item Difficulty Analyses for C1

	Rates of instructors' assessment of item difficulty perception	Item difficulty
Item 1	62% Low 34% Moderate 4% High	0.95
Item2	64% Low 30% Moderate 6% High	0.86
Item3	26% Low 56% Moderate 18% High	0.90
Item4	10% Low 68% Moderate 22% High	0.80
Item5	12% Low 66% Moderate 22% High	0.85
Item6	68% Low 30% Moderate 2% High	0.98
Item7	32% Low 54% Moderate 14% High	0.92
Item8	44% Low 38% Moderate 18% High	0.86
Item9	20% Low 48% Moderate 32% High	0.83
Item10	36% Low 56% Moderate 8% High	0.86

The most striking data in Table 12 highlights that Item 3, which had a very low difficulty value, was accepted by 56% of instructors at a moderate level. It was previously reported the discrimination index of Item 3 was also outside the acceptable value for item analysis. A similar difference was observed in Items 4, 5, 7, 9, and 10. In Item 8, although the percentages were close, 38% of instructors thought the item was of moderate difficulty for an item with very low difficulty level, and as a result, should be considered easy. The rate of instructors who found the difficulty level low was 44%. In short, instructors almost equally chose easy and moderate difficulty levels for this item. Additionally, data in the C1 test clearly showed instructors' item difficulty perceptions were not compatible with the item analyses. In other tests, this discrepancy was lower than in the C1 test, however, it can still be seen instructors' perceptions of item difficulty values were mostly inconsistent with the analyses.

Previous research also indicates significant gaps in the assessment and evaluation competencies of instructors. According to Ustabulut (2021) as well as Erdoğdu and Kurt (2012), instructors' ability to analyse exam results and make evaluations aligned with learning objectives is at a moderate level. On the other hand, Yıldız and Tepeli (2014) found instructors demonstrate high competency in applying contemporary assessment and evaluation methods. However, studies by Hatipoğlu (2015), Mede and Atay (2017), and Ölmezer-Öztürk and Aydın (2018) revealed instructors lack sufficient knowledge in the field of assessment and evaluation. Similar findings are observed in international studies with Ahmadi and Ketabi (2020), Bahtiar and Purnawarman (2020), Fitriyah et al. (2022), and Latif (2021) reporting instructors feel inadequate in the area of foreign language assessment literacy. Additionally, Bøhn and Tsagari (2021), Firoozi et al. (2019), Liu and Li (2020), Razavipour and Rezagah (2018), and Sultana (2019) emphasized instructors lack necessary skills in exam preparation and learner evaluation.

5.3. Analyses for instructors' predictions

In the final study stage, agreement analysis was conducted between the instructors' item difficulty perceptions and item difficulty analyses. By reaching the quantitative results for evaluations made in both stages with agreement analysis, the rate of difference between instructors' item difficulty perceptions and item analyses was determined.

Percentages of agreement determined after applying the formula are presented in Table 13.

Table 13

Instructors' Item Difficulty Perception and Item Difficulty Analysis Compatibility

Test	Agreement rate	Agreement value
B1.1	40%	Low
B1.2	40%	Low
B2.1	40%	Low
B2.2	80%	High
C1	40%	Low

As can be seen in Table 13, instructors' item difficulty perceptions in general along with difficulty values in the item analyses were not consistent. Thus, the percentage agreement for each instructor was calculated to determine their success rate in terms of predicting the difficulty level of an item. As a result, the percentage of correctly predicted items for each instructor was calculated.

The percentage of correct predictions by each instructor ranged from 17.5% to 72.5%. Variation in success rates indicated a significant disparity in instructors' ability to accurately predict item difficulty. For example, some instructors were able to match actual difficulty levels of items more accurately than others. Instructors with success rates above 50% showed relatively high predictive accuracy. Therefore, these instructors might possess better intuition or experience in accurately judging item difficulty. For instance, instructors with success rates around 72.5% might be leveraging their extensive experience and/or specific training in assessment. Whereas instructors with success rates below 30% struggled to accurately predict item difficulty. This could be due to various factors, such as less experience, lack of specific training in item difficulty assessment, and/or differing perceptions of what constitutes item difficulty.

The next step involved development of a confusion matrix to understand the type of errors made by each instructor. A confusion matrix allows for comprehensive evaluation of how well a model performs as well as where it might go wrong (Witten et al., 2005). In our study, in the context of evaluating instructors' predictions of item difficulty, using a confusion matrix aided in identifying specific patterns regarding their predictions. The matrix included:

- True Positives (TP): Correctly predicted difficulty levels.
- False Positives (FP): Predicted a higher difficulty level than actual.
- False Negatives (FN): Predicted a lower difficulty level than actual.

For instance, a moderate item predicted as moderate was a true positive. A false positive could be exemplified by a moderate item predicted as easy. Whereas a moderate item predicted as difficult was an example of a false negative. Considering Instructor 1 as an example, for easy items, this instructor correctly identified 18 as easy, but incorrectly classified eight as medium and one as difficult. For medium items, Instructor 1 correctly predicted five as medium, but mistakenly labelled six as easy and one as difficult. When it came to difficult items, Instructor 1 correctly identified none, misclassifying one as easy and none as medium.

Table 14

Confusion matrix sample (for Instructor 1)

Predicted	**Actual**	**Count**	**Instructor ID**
Easy	Easy	18	1
Easy	Moderate	6	1
Easy	Difficult	1	1
Moderate	Easy	8	1
Moderate	Moderate	5	1
Moderate	Difficult	0	1
Difficult	Easy	1	1
Difficult	Moderate	1	1
Difficult	Difficult	0	1

Next, to understand common patterns and types of errors made by all instructors, we aggregated the confusion matrices for each instructor. The aggregated confusion matrix revealed several key insights. For easy items, 37.18% were correctly predicted as easy, 48.44% were incorrectly predicted as medium, and 14.38% incorrectly predicted as difficult. For medium items, 46.90% were correctly predicted as medium, while 31.54% were incorrectly predicted as easy, and 21.57% incorrectly predicted as difficult. For difficult items, only 33.33% were correctly predicted as difficult, with 50.98% being incorrectly predicted as medium, and 15.69% as easy.[1]

The most frequent misclassification occurred between easy and medium items. For example, instructors tended to overestimate the difficulty of easy items, predicting them as medium, and underestimating the difficulty of medium items, predicting them as easy. This indicated a general tendency to perceive items within a narrower range of difficulty, often defaulting to a medium rating.

Moreover, difficult items were often underestimated, with a significant portion being predicted as medium and some even as easy. These suggested instructors had a challenging time accurately identifying items as difficult, likely due to a lack of clear distinguishing characteristics for such items.

[1] Success percentages for each instructor are presented in Appendix 1.

Medium items were somewhat more accurately predicted compared to easy and difficult items, with 46.90% correct predictions. However, there was still notable confusion, especially with predictions leaning towards easy.

The fact the prediction agreement percentages of instructors were low in general, brought to mind the question of whether experience, program graduation, education level, and so forth had some positive effect on item difficulty perception.

Instructors' predictions of item difficulty were compared to actual difficulty levels to calculate success rates. Prediction success rates were calculated as the percentage agreement between instructors' predictions and actual difficulty levels. To understand which demographic variables affected these success rates, group differences were analysed through independent t-tests and ANOVA (based on the number of variable categories) for each CEFR level separately (B1, B2, and C1). Descriptive statistics for instructors' success rates at different CEFR levels are presented in Table 15.

Table 15

Instructors' Success Rates at Different CEFR Levels

CEFR Level	N	Mean	Std. Deviation
B1	40	0.512	0.129
B2	40	0.458	0.136
C1	40	0.372	0.147

Assessment training was determined to be the only binary variable with significant group difference in instructors' prediction success rate at the B2 level. The t-test results also showed a significant difference (t(20.94) = -2.682, p = 0.0131), indicating instructors with assessment training had significantly higher success rates.

To investigate the impact of various demographic variables on instructors' success rates in predicting item difficulty at different CEFR levels, a series of ANOVA tests were conducted. The variables examined included education level, education department, and teaching experience. Multifactorial ANOVA was also performed to explore possible interactions which effected item difficulty prediction success rates for instructors, but no significant interaction was found.

Furthermore, ANOVA results for the B1 level indicated none of the demographic variables examined (i.e., education, department, experience) had a significant effect on instructors' success rates. The lack of significant findings suggested these factors did not influence instructors' ability to predict item difficulty at the B1 level.

Similarly, for the B2 level, ANOVA results showed no significant effect of demographic variables (i.e., education, department, experience) on instructors' success rates. This indicated these factors did not significantly impact instructors' accuracy in predicting item difficulty at the B2 level.

Contrary to expectations, ANOVA analysis for the B1 and B2 levels did not reveal any significant effects of the demographic variables examined. This suggested factors such as education level, department, and teaching experience did not substantially influence instructors' ability to predict item difficulty for lower and intermediate proficiency levels.

For the C1 level, ANOVA results revealed a significant effect of 'Experience' on success rates (F(3, 38) = 3.7233, p = 0.0193). This suggested teaching experience played a crucial role in predicting the difficulty of C1 level items. To further explore specific group differences within the 'Experience' variable, post-hoc comparisons using the Tukey HSD test were performed. The Tukey HSD test results indicated instructors with more than 10 years of experience had significantly higher success rates compared to those with 2-4 years of experience. Considering almost all of the items at the C1 level were classified as easy according to the item analysis results, it can be concluded this achievement difference was not important. Because at this point, guessing factors can be considered high. Moreover, the fact there was no significant effect of experience on prediction success at other levels supported this finding.

Therefore, our study's findings indicated that demographic factors such as education level, education department, and teaching experience did not significantly affect instructors' ability to accurately predict item difficulty at the B1 and B2 CEFR levels, aligning with Tao's (2014) assertion that formal education, including undergraduate and postgraduate degrees, may not substantially enhance teachers' competencies in language assessment. This highlighted the inadequacy of short-term academic courses, as noted by Sultana (2019) as well as Yan and Fan (2021) in fostering a deep understanding of assessment. The current study also reinforced Tao's (2014) argument that teaching experience alone is insufficient for developing assessment literacy, also supporting Fitriyah, Massitoh, and Widiati (2022) which called for ongoing and targeted professional development. While this study determined teaching experience had a significant impact on predicting item difficulty at the C1 level, particularly for those with over 10 years of experience, the influence of guessing due to the ease of items should be considered. This was consistent with Levi and Inbar-Lourie (2020) who concluded theoretical knowledge without practical application may not lead to better performance. Similarly, Ölmezer-Öztürk and Aydın (2018) found even experienced instructors require structured, experiential learning opportunities to improve their assessment skills.

5. Conclusion

This study highlighted critical gaps in instructors' ability to predict item difficulty within the context of teaching Turkish as a foreign language. Despite the availability of research regarding the competencies of educators, there remained a scarcity of descriptive studies focused on identifying specific deficiencies, especially in the realm of assessment and evaluation. This study aimed to fill this gap by investigating the accuracy of instructors' perceptions of item difficulty along with the implications for validity and reliability of exams.

Our findings revealed instructors often struggled with correctly assessing item difficulty, which had serious implications on test validity. This result aligned with previous research by Alderson (1993) and Bachman (2002), who both observed significant challenges among even experienced educators in predicting the difficulty levels of test items. These studies further demonstrated instructors, particularly those with extensive experience in exam preparation, may overestimate their ability to accurately gauge item difficulty. In fact, more experienced instructors tended to have lower agreement percentages in their item difficulty predictions. Similarly, Shohamy (1984), observed response format familiarity, particularly with multiple-choice items, can lead to misjudgements regarding item difficulty. This suggested experience alone is insufficient for accurate item difficulty assessment without necessary training and established analytical skills.

Moreover, correct determination of item difficulty is crucial for ensuring items are appropriately classified according to a learners' language proficiency level. Inaccuracies in this process can compromise both exam validity and reliability. For instance, instructors who misclassify difficult items as "easy" raise questions about their ability to create level-appropriate assessments. This issue is further supported by Bejar (1983), who demonstrated the use of anchor-based methods and structured training can significantly enhance accuracy of item difficulty predictions. As a result, these findings underscored the need for more comprehensive training programs focused on item analysis and test construction.

Furthermore, this study drew attention to the need for improving instructors' understanding of item characteristics, particularly those which influence difficulty levels. Training focused on recognizing features of challenging items can help instructors more accurately predict item difficulty, leading to improved item classification, and ultimately more valid and reliable assessments. Freedle and Kostin (1999) emphasized topic familiarity plays a significant role in item difficulty predictions, while Rupp, Garcia, and Jamieson (2001) highlighted the importance of lexical knowledge. Incorporating these elements into professional development programs can greatly benefit instructors’ assessment literacy.

Therefore, within the scope of our study, the following suggestions were made:

-Teachers should undergo specialized training aimed at enhancing their proficiency in item construction, with particular emphasis on developing items that measure higher-order cognitive skills.

-Teacher education programs at universities should incorporate a greater number of courses regarding assessment and evaluation, supplemented with practical, hands-on learning opportunities.

-Mentorship and support systems should be established, whereby experienced educators provide guidance and feedback to less experienced teachers, particularly in the context of test item development.

-The process of item construction should be standardized through the implementation of structured guidelines and templates, and all test items should be subject to pilot testing prior to formal administration.

-Regular reviews and revisions of test items should be conducted, particularly for items with low discrimination indices, to ensure assessment reliability and validity.

-Mechanisms for structured feedback should be developed to improve teachers' accuracy in predicting item difficulty, along with more advanced training focused on item analysis and difficulty estimation.

-In high-stakes examinations, particularly at the C1 level and above, a stronger focus should be placed on the development of test items which assess advanced cognitive abilities, in alignment with Bloom's Revised Taxonomy.

-The implementation of mandatory pilot testing and comprehensive data analysis for all assessments should be enforced to ensure the validity and discriminatory power of test items prior to their final administration.

References

Ahmadi, M. R. S., & Ketabi, S. (2020). Features of language assessment literacy in Iranian English language teachers' perceptions and practices. *Journal of Teaching Language Skills*, *38*(1), 191-223. https://doi.org/10.22099/jtls.2020.34843.2739

Alderson, C. (1993). Judgments in language testing. In D. Douglas & C. Chapelle (Eds.), *A new decade in language testing* (pp. 46-57). TESOL.

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

Altıntaş, N. (2022). *Yabancı dil olarak Türkçe öğreten öğretim elemanlarının sınav hazırlama ve ölçme değerlendirme yeterlik algıları [Perceptions of Exam Preparation and Assessment Competencies Among Instructors Teaching Turkish as a Foreign Language]* (Thesis Number: 747926) [Master Thesis, Dokuz Eylül University]. Turkish Council of Higher Education Thesis Center.

Aydemir, P. D. Y., & Çiftçi, Y. Ö. (2008). A study on the questioning skills of literature teacher candidates. *Van Yüzüncü Yıl University Journal of Education Faculty*, *5*(2), 103-115. Retrieved from https://dergipark.org.tr/en/pub/yyuefd/issue/13714/166035

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing*, *19*(4), 453-476. https://doi.org/10.1191/0265532202lt240oa

Bahtiar, I., & Purnawarman, P. (2020). Investigating English teachers' comprehension in language assessment literacy (LAL). *Advances in Social Science, Education and Humanities Research*, *508*, 303-310. https://doi.org/10.2991/assehr.k.201214.253

Bejar, I. I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement*, *7*(3), 303-310. https://doi.org/10.1177/014662168300700305

Bøhn, H., & Tsagari, D. (2021). Teacher educators' conceptions of language assessment literacy in Norway. *Journal of Language Teaching and Research*, *12*(2), 222-233. https://doi.org/10.17507/jltr.1202.02

Boylu, E. (2019). *Yabancılara Türkçe öğretiminde ölçme değerlendirme uygulamaları ve standart oluşturma. [Measurement and evaluation practices and standardization in teaching Turkish to foreigners]* (Thesis Number: 542435) [Doctoral Dissertation, Çanakkale Onsekiz Mart University]. Turkish Council of Higher Education Thesis Center.

Çavuşoğlu, R., & Işık, A. D. (2021). Assessment and evaluation process of Turkish language teaching centers (TÖMER). *The Journal of Limitless Education and Research*, 6(2), 291-315. https://doi.org/10.29250/sead.958711

Cohen, L., Manion, L., & Morrison, K. (2000). *Research methods in education (5th ed.)*. Routledge.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Cengage Learning.

Erdoğdu, M. Y., & Kurt, F. (2012). Investigation of teachers' perception of their competencies in assessment and evaluation in terms of certain variables. *Electronic Journal of Education Sciences*, *1*(2), 23-36. Retrieved from https://dergipark.org.tr/en/pub/ejedus/issue/15938/167586

Firoozi, T., Razavipour, K., & Ahmadi, A. (2019). The language assessment literacy needs of Iranian EFL teachers with a focus on reformed assessment policies. *Language Testing in Asia*, *9*(1), 1-12. https://doi.org/10.1186/s40468-019-0078-7

Fergadiotis, G., Swiderski, A., & Hula, W. D. (2019). Predicting confrontation naming item difficulty. Aphasiology, *33*(6), 689–709. https://doi.org/10.1080/02687038.2018.1495310

Fitriyah, I., Massitoh, F., & Widiati, U. (2022). Classroom-based language assessment literacy and professional development need between novice and experienced EFL teachers. *Indonesian Journal of Applied Linguistics*, *12*(1), 124-134. https://doi.org/10.17509/ijal.v12i1.46539

Freedle, R., & Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? The case for construct validity of the TOEFL's minitalks. *Language Testing*, *16*(1), 2-32. https://doi.org/10.1177/026553229901600101

Gedik, E. (2017). *Yabancı dil olarak Türkçe öğretiminde ölçme ve değerlendirme [Assessment and Evaluation in Teaching Turkish as a Foreign Language]* (Thesis Number: 458489) [Master Thesis, Istanbul Arel University]. Turkish Council of Higher Education Thesis Center.

Hambleton, R. K., & Jirka, S. J. (2006). Anchor-based methods for judgmentally estimating item statistics. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 399-420). Lawrence Erlbaum Associates.

Hatipoğlu, Ç. (2015). Language assessment in English teacher education programs in Turkey. *Journal of Language Testing*, *32*(2), 243-265. https://doi.org/10.1177/0265532214565635

Işıkoğlu, M. (2015). *Yabancı dil olarak Türkçe öğretiminde kullanılan yeterlik sınavlarının madde yazımı bakımından incelenmesi (Mersin ve Sakarya üniversiteleri örneği) [Analysis of proficiency exams developed for teaching Turkish as a foreign language in terms of item writing: Samples of Mersin and Sakarya Universities]* (Thesis Number: 394791) [Doctoral Dissertation, Gazi University]. Turkish Council of Higher Education Thesis Center.

Karagöl, E. (2020). Proficiency exams in teaching Turkish as a foreign language in TÖMER (Turkish and foreign languages research and application centers). *Journal of Language and Linguistic Studies*, *16*(2), 930-947. https://doi.org/10.17263/jlls.759347

Kostin, I. (2004). Exploring item characteristics that are related to the difficulty of TOEFL dialogue items. *ETS Research Report Series*, *2004*(1), i-71. https://doi.org/10.1002/j.2333-8504.2004.tb01938.x

Kutlu, Ö., Doğan, C. D., & Karakaya, İ. (2010). *Öğrenci başarısının belirlenmesi: Performansa ve portfolyoya dayalı durum belirleme [Assessing Student Achievement: Performance and Portfolio-Based Evaluation] (3rd ed.)*. Pegem Akademi.

Latif, M. W. (2021). Exploring tertiary the EFL practitioners' knowledge base component of assessment literacy: Implications for teacher professional development. *Language Testing in Asia*, *11*(1), 1-22. https://doi.org/10.1186/s40468-021-00130-9

Levi, T., & Inbar, O. (2019). Assessment literacy or language assessment literacy: Learning from the teachers. *Language Assessment Quarterly*, *17*(3), 1-15. https://doi.org/10.1080/15434303.2019.1692347

Liu, J., & Li, X. (2020). Assessing young English learners: Language assessment literacy of Chinese primary school English teachers. *International Journal of TESOL Studies*, *2*, 36-49. https://doi.org/10.46451/ijts.2020.12.05

Mercan, Ö., & Göktaş, B. (2023). Türkçenin yabancı dil olarak öğretimi sertifika programları ile CELTA'nın karşılaştırılması: Bir program önerisi [A Comparison of Turkish as a Foreign

Language Certification Programs and CELTA: A Program Proposal]. *Cumhuriyet International Journal of Education*, *12*(3), 715-731. https://doi.org/10.30703/cije.1258699

Mede, E., & Atay, D. (2017). English language teachers' assessment literacy: The Turkish context. *Dil Dergisi*, *168*(1), 1-5. Retrieved from https://www.ceeol.com/search/article-detail?id=548906

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded source book (2nd ed.)*. Sage Publications.

Nichols, T. R., Wisner, P. M., Cripe, G., & Gulabchand, L. (2010). Putting the kappa statistic to use. *The Quality Assurance Journal*, *13*(3-4), 57-61. https://doi.org/10.1002/qaj.481

Oktay, M. R. (2015). *An analysis of sub-questions in Turkish textbooks used in teaching Turkish as a foreign language in terms of cognitive levels in Bloom's Taxonomy* (Thesis Number: 463400) [Master Thesis, Başkent University]. Turkish Council of Higher Education Thesis Center.

Ölmezer Öztürk, E., & Aydın, B. (2018). Developing and validating language assessment knowledge scale (LAKS) and exploring the assessment knowledge of EFL teachers. *Hacettepe University Journal of Education*, *34*(3), 602-620. https://doi.org/10.16986/HUJE.2018043465

Özcan, S., & Akçan, K. (2010). Investigation of questions prepared by science teacher candidates in terms of content and Bloom's Taxonomy. *Kastamonu Journal of Education*, *18*(1), 323-330. Retrieved from https://dergipark.org.tr/en/pub/kefdergi/issue/49066/626064

Özdemir, S. (2023). Yabancı dil olarak Türkçe öğretiminde ölçme ve değerlendirmeye yönelik araştırmaların eğilimleri [Trends in Research on Assessment and Evaluation in Teaching Turkish as a Foreign Language]. *İnönü Üniversitesi Eğitim Fakültesi Dergisi*, *24*(1), 537-559. https://doi.org/10.17679/inuefd.1071671

Özdemir, S., & Eke, H. (2023). Yabancılara Türkçe öğretiminde uygulanan A1 ve A2 kur sınavlarının madde yazma ilkeleri açısından incelenmesi [An Examination of A1 and A2 Level Exams in Teaching Turkish to Foreigners in Terms of Item Writing Principles]. *Trakya Eğitim Dergisi*, *13*(1), 365-380. https://doi.org/10.24315/tred.700445

Razavipour, K., & Rezagah, K. (2018). Language assessment in the new English curriculum in Iran: Managerial, institutional, and professional barriers. *Language Testing in Asia*, *8*(1), 1-12. https://doi.org/10.1186/s40468-018-0061-8

Rupp, A. A., Garcia, P., & Jamieson, J. (2001). Combining multiple regression and CART to understand item difficulty in second language reading and listening comprehension tests. *International Journal of Testing*, *1*(3-4), 185-216. https://doi.org/10.1080/15305058.2001.9669479

Sertdemir, E. (2021). *An analysis of Turkish proficiency exams used in teaching Turkish as a foreign language according to the Revised Bloom's Taxonomy* (Thesis Number: 652111) [Master Thesis, Çanakkale Onsekiz Mart University]. Turkish Council of Higher Education Thesis Center.

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing*, *1*(2), 147-170. https://doi.org/10.1177/026553228400100203

Şimşek, A. (2016). A comparative analysis of common mistakes in achievement tests prepared by school teachers and corporate trainers. *European Journal of Science and Mathematics Education*, *4*(4), 477-489. https://doi.org/10.30964/ejsme.v4i4.467

Tao, N. (2014). *Development and validation of classroom assessment literacy scales: English as a foreign language (EFL) teachers in a Cambodian higher education setting* (Unpublished doctoral dissertation). Victoria University, Australia. Retrieved from https://vuir.vu.edu.au/25850/

Ustabulut, M. Y. (2021). Examination of education program literacy levels of instructors teaching Turkish as a foreign language. *Fırat University Journal of Social Sciences*, *31*(3), 1235-1243. https://doi.org/10.18069/firatsbed.941957

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2005). Practical machine learning tools and techniques. In *Data mining* (2nd ed., Vol. 2, pp. 403-413). Elsevier.

Yıldız, Ü., & Tepeli, Y. (2014). A study on teacher qualifications in teaching Turkish as a foreign language. *International Journal of Language Academy*, *2*(4), 564-578. https://doi.org/10.18033/ijla.182

Article Information Form

Authors Contributions: All authors contributed equally to the writing of this paper. All authors read and approved the final manuscript.

Conflict of Interest Disclosure: No potential conflict of interest was declared by authors.

Copyright Statement: Authors own the copyright of their work published in the journal and their work is published under the CC BY-NC 4.0 license.

Supporting/Supporting Organizations: No grants were received from any public, private or non-profit organizations for this research.

Ethical Approval: It is declared that during the preparation process of this study, scientific and ethical principles were followed and all the studies benefited from are stated in the bibliography. The ethics committee report for this study was obtained from the Istanbul University Rectorate Social and Human Sciences Research Ethics Committee with the decision dated 11.01.2024 and numbered 2024/19.

Plagiarism Statement: This article has been scanned by iThenticate.

Appendix 1

Instructor ID	Success Percentage
1	57.5
2	52.5
3	35
4	22.5
5	32.5
6	50
7	30
8	25
9	42.5
10	40
11	32.5
12	50
13	25
14	52.5
15	35
16	35

Appendix 1 (Continued)

17	27.5
18	55
19	25
20	37.5
21	42.5
22	32.5
23	17.5
24	40
25	25
26	55
27	52.5
28	52.5
29	25
30	30
31	42.5
32	72.5
33	40
34	47.5
35	32.5
36	30
37	40
38	37.5
39	50
40	50
41	47.5
42	57.5
43	65
44	42.5
45	375
46	35
47	50
48	475
49	175
50	40
51	22.5