Education and Science

Vol 49 (2024) No 220 239-257

Original Article

Validating Tests of Turkish L2 Receptive Skills: An Argument-Based Validation Study through Standard Setting [*]

Yiğit Savuran [1], Zühal Çubukçu [2]

Abstract

This paper reports on a validation study based on an assessment-use argument for four level tests developed under a larger project for adult learners of Turkish as a second language (L2). We treat the test scores as data on which we build the validity claim that the tests accurately classify learners into the intended Common European Framework of Reference for Languages (CEFR) levels. Four level tests (A1, A2, B1, and B2), each comprising listening and reading tasks, were administered to mixed groups of students, including those at Pre-A1 and C1 levels. Cut scores for each of the four tests were determined through the Angoff method and were considered as backing for the validity claim. To provide warrants, we assumed a 50% probability of a test-taker being at a CEFR level for Chi-square goodness-of-fit tests, which were conducted to assess the statistical significance between the expected and observed numbers of students under and above the cut score for each level. The distribution of student scores—with acceptable item difficulty and discrimination indices—cut scores placed in the intervals between adjacent levels, and chi-square analyses of all four tests enabled us to conclude that the tests have the potential to validly demonstrate the intended learner performance. With its innovative design and techniques in data collection and analysis, this paper offers theoretical, methodological and practical insights for practitioners in Turkish L2, based on solid empirical evidence.

Keywords

Turkish L2
Test validation
Assessment-use argument
Standard-setting

Article Info

Received: 06.02.2023
Accepted: 06.12.2024
Published Online: 10.27.2024

DOI: 10.15390/EB.2024.12959

Introduction

There is a growing demand among international students to study at higher education levels in Türkiye (Council of Higher Education, 2023). Most of these students are grantees of the Türkiye Scholarships program. The scholarship offers a comprehensive list of benefits, including university and department placements, tuition fees, accommodation, flight tickets, health insurance, a monthly stipend, as well as a Turkish language course, which makes it attractive to international students. The

[*] This article is derived from Yiğit Savuran's PhD dissertation entitled "Developing basic and independent level descriptors for Turkish as a foreign language", conducted under the supervision of Zühal Çubukçu.
One of the authors of the article, Zühal Çubukçu, sadly passed away during the review process. The article has been dedicated by the other author in her honor, commemorating her legacy.

[1] University of Florida, College of Liberal Arts and Sciences, Department of Linguistics, USA; Anadolu University, School of Foreign Languages, Türkiye, yigitsavuran@gmail.com

[2] Eskişehir Osmangazi University, Faculty of Education, Department of Educational Sciences, Türkiye, zuhal_cubukcu@hotmail.com

program accommodates around 15,000 students per year and has more than 150,000 alumni to date (Türkiye Bursları, n.d.). Additionally, for several reasons, such as recent political tensions, socio-economic conflicts, and mobility programs offered by various institutions like Erasmus+ (European Commission, n.d.) and Study in Türkiye (Council of Higher Education, 2024), an increasing number of students prefer Türkiye for their higher education. The students, whether scholarship grantees or those studying with their own funds, usually enroll in Turkish Language Centers scattered around the country to learn Turkish through a one-year intensive Turkish language (Turkish L2) instruction program.

The Turkish Language Centers in Türkiye generally follow the Common European Framework of Reference for Languages (CEFR) guidelines and terminology in designing their curricula and assessment practices. The scholarship students, for instance, must successfully complete the CEFR B2 or C1 level as part of their program. This requirement and the increasing number of students learning Turkish L2 create an opportunity for researchers to develop well-structured curricula and reliable, valid assessment tools based on CEFR principles. To develop such tools, as stated in CEFR (Council of Europe [CoE], 2001, 2020), researchers need to have locally designed and adapted descriptors that demonstrate learner performance at various levels. This need prompted us to carry out the larger study, which focused on the development of such descriptors, together with the implementation of level tests for receptive skills and tasks for productive ones, addressing the needs of learners of Turkish as a second language.

The larger study aimed to develop descriptors for the assessment needs of Turkish L2 learners, with a specific focus on higher education levels, in four phases: 'preparation', 'development', 'implementation', and 'validation'. In the preparation phase, we first worked on adapting descriptors. The development stage involved selecting appropriate descriptors for assessment purposes. During implementation, students responded to level tests and tasks, which was followed by determining cut scores for each for the standard-setting study in the validation phase.

The preparation and development phases were reported in detail in our previous study (Savuran & Çubukçu, 2021). The present study, however, aims to provide empirical validation of the tests administered in the implementation phase and to report on the standard-setting process conducted in the validation phase through the assessment-use argument (Bachman & Palmer, 2010; Papageorgiou & Tannenbaum, 2016), and the assumption of a 50% probability of being at a CEFR level (De Jong & Benigno, 2017; Harsch, 2019; North, 2000).

Two types of skills in language testing—receptive and productive—have some differences in the methodology applied in the validation process, especially within argument-based validation due to the types of items used. Most tests addressing receptive skills (e.g., listening and reading) deploy multiple-choice items, while those assessing productive skills (e.g., writing and speaking) require test takers to provide open-ended responses. With that said, to collect a validity argument for a test (Kane, 2006, 2013), a researcher needs to analyze different elements depending on the skill being assessed. In receptive skills, it is mostly the test itself, the test scores, and the test-taker's choices for each item that form evidence for building a validity argument regarding the test's use and the interpretation of the scores. On the other hand, for the validity evidence of the tests for productive skills, a researcher examines the validity of the rating scale as well as the item and test-takers' open-ended responses to that item (Chapelle & Voss, 2014). With this in mind, we excluded the validation of productive skills, writing and speaking, since they require a different approach to validation; this is the subject of another study. The current study, therefore, focuses on validating the tests of receptive skills (namely listening and reading) at four CEFR levels (A1-B2) through assessment-use argument and standard setting.

Theoretical Background: Argument-Based Validation

Validity has long been discussed in the educational measurement literature (Cizek, 2012; Cureton, 1951; Messick, 1989), and recently it has been acknowledged that it is the interpretations of the test scores and their proposed uses that can be validated, not the test itself (American Educational Research Association [AERA], 2014; Bachman, 2005; Fulcher, 2015; Kane, 1994). Based on such consensus, Chapelle and Voss (2014) suggest four validation approaches, among which 'evidence gathering' and 'argument-based' have recently gained prominence (Cheng & Sun, 2015).

The argument-based approach to validation has sparked considerable interest among educational measurement researchers, as it provides a practical guideline for building argumentation based on the test context and uses and interpretations of the scores (Bachman, 2005; Chapelle, Enright, & Jamieson, 2010; Kane, 2013; Knoch & Chapelle, 2018). To build a validity argument, researchers are expected to gather validity evidence supporting the proposed uses and interpretations of the test scores (Chapelle, Enright, & Jamieson, 2008; Im, Shin, & Cheng, 2019); however, key components of the validation process are the logic and argumentation, rather than just the evidence itself (Lavery, Bostic, Kruse, Krupa, & Carney, 2020).

The argument-based approach utilizes two types of arguments for different purposes: Interpretation/Use Argument (IUA) and Validity Argument (Kane, 2006, 2013). Independent of the argument type, Kane's approach provides practical ways of validation such as how to plan the interpretation of the scores, conduct research, organize research results for validity argumentation, and challenge the validity argument (Chapelle et al., 2010). Building on Kane's work and in relation to the validity argument type, Bachman and Palmer (2010) presented the Assessment Use Argument (henceforth, AUA), which underscores the utilization of tests by incorporating Messick's (1989) unified model, a model upon which Kane's approach is constructed (Kane, 2013).

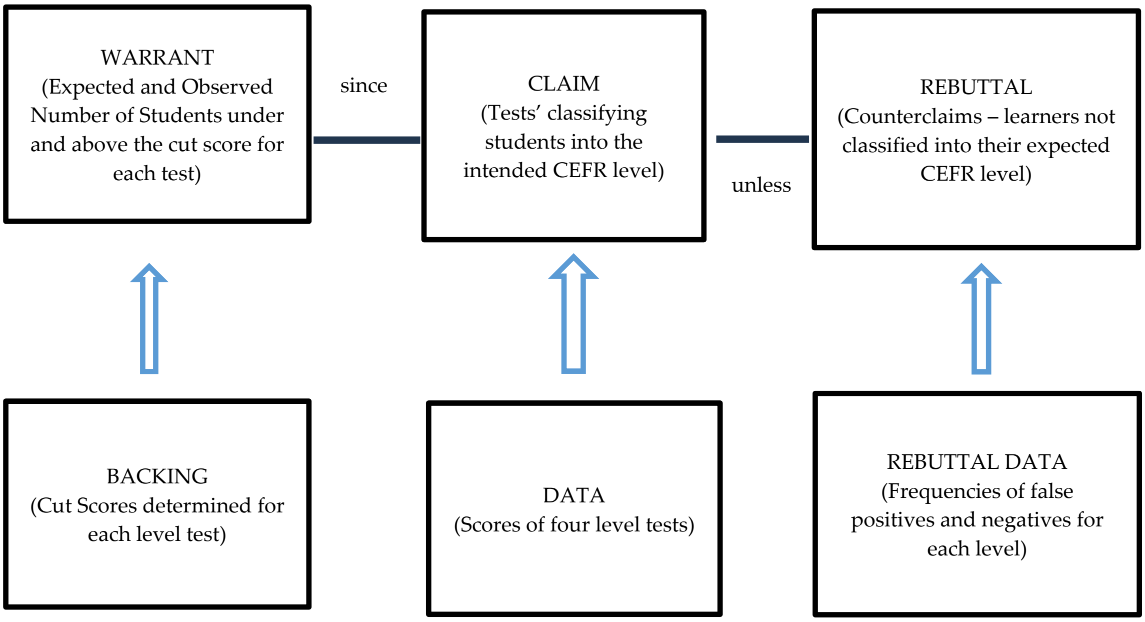
WARRANT
(Expected and Observed Number of Students under and above the cut score for each test)
since
CLAIM
(Tests' classifying students into the intended CEFR level)
unless
REBUTTAL
(Counterclaims – learners not classified into their expected CEFR level)
BACKING
(Cut Scores determined for each level test)
DATA
(Scores of four level tests)
REBUTTAL DATA
(Frequencies of false positives and negatives for each level)

Figure 1. Assessment Use Argument Structure (adapted from Papageorgiou & Tannenbaum, 2016)

The AUA framework suggests generating claims, as illustrated in Figure 1, which are rooted in and supported by data. As Papageorgiou and Tannenbaum (2016) articulate, to provide claims, "a warrant is stated, which in turn is supported by backings arising from theoretical or empirical evidence" (p. 112). Within this framework, rebuttals can serve as 'counterclaims' and may challenge the initial claims if substantiated with sufficient data (Bachman, 2005; Papageorgiou & Tannenbaum, 2016). Drawing upon Bachman and Palmer (2010), Papageorgiou and Tannenbaum (2016) present four claims that can be integrated into an AUA study. Claim 1 contends that the use of assessments and the ensuing decisions bring about beneficial consequences for stakeholders. Claim 2 asserts that the "decisions made on the basis of assessment-driven interpretations take community values, pertinent legal stipulations into account, and are equitable for those stakeholders impacted by such decisions" (Bachman & Palmer, 2010, p. 111). Claim 3 emphasizes that the interpretation of abilities being assessed should correspond with the curriculum, theory, and the target language use domain. Claim 4 relates to the consistency of assessment outcomes, manifest as numeric scores or performance descriptions, referred to as 'assessment records'. Although it is posited that these claims begin with the consequences of the tests and are inferentially connected, our focus is on Claims 3 and 4, regarding the interpretation of test-taker abilities and the consistency of assessment results, respectively (Papageorgiou & Tannenbaum, 2016).

Building on the AUA framework outlined above, the current study considers the level tests of Turkish L2 receptive skills as the 'data' upon which the 'claims' (i.e., the classification of learners into the intended CEFR level by test scores) are formulated. To this end, mixed student groups participated in four level tests within physical classroom settings. For instance, the B1 level test was administered to A2, B1, and B2 students to facilitate an in-depth analysis of learner performance by level and, hence, provide 'data.' Score averages from students at different levels were compared with cut scores established via the Angoff standard setting method, strengthening the inference from data to claim. To lend further support to these comparisons and furnish warrant for the validity claim, we adopted the notion of a 50% probability for accurately solving items at a given level, as discussed in the literature (De Jong & Benigno, 2017; Harsch, 2019), for chi-square goodness-of-fit tests. The chi-square analyses sought to determine if there was statistical significance in the differences between observed and expected numbers of students taking the same level test. Counterclaims that argue the tests do not accurately classify students into the intended levels were treated as 'rebuttals,' underpinned by 'rebuttal data' which elucidate the disparities between observed and expected figures (i.e., false positives and false negatives).

What it means to be at a CEFR level: The assumption of 50% Probability

Unlike the natural sciences, where assessments are based on tangible facts, the constructs measured in language proficiency are abstract, posing significant challenges for language assessors. Douglas's (2010) 'rubber ruler' metaphor, symbolizing potential variances in measuring units' nature, underlines the necessity of precision in language proficiency calibration. Such precision could be attained when latent language proficiency characteristics become salient features, buttressed by both quantitative and qualitative facets, akin to North's (2000) work that underpins the CEFR's illustrative scales and descriptors. According to North (2000), the study presumes that a learner, at the commencement of a level, has a 50% probability of resolving items and tasks at that level, delineating level boundaries based on this probability model. This supposition indicates that as tasks simplify, a learner's likelihood of successfully solving items increases and vice versa (Harsch, 2019). Similar probability values have been reported in studies developing proficiency frameworks other than CEFR. Zwick, Senturk, Wang, and Loomis (2001) proposed an interval between .50 and .80 for item anchoring and matching. Likewise, Gomez, Noah, Schedl, Wright, and Yolkut (2007) suggested that an item might signify learner aptitude if over 50% of examinees at a particular level correctly respond to it and fewer than 50% at a lower level do so. In their alignment study of the GSE with other scales, De Jong and Benigno (2017) state that "… being at B1 implies an expectation to successfully perform 50% of all tasks at B1, or to possess a 50% chance of successfully performing any given task at B1" (p. 5). Table 1 exhibits their postulated probabilities for different level learners at various CEFR levels, adapted from North (2000).

Table 1. Probability Estimates for Learners at Different CEFR Levels (adapted from De Jong & Benigno, 2016)

		Learners at Level				
		A1	**A2**	**B1**	**B2**	**C1**
Descriptors/ Tasks at Level	**C1**	.00	.00	.03	.17	**.50**
	B2	.00	.02	.12	**.50**	.83
	B1	.03	.12	**.50**	.88	.97
	A2	.18	**.50**	.88	.98	1.00
	A1	**.50**	.82	.97	1.00	1.00

The 50% probability threshold should not be regarded as inflexible, since it is grounded in theoretical considerations, and it is the role of test developers to define and declare the knowledge, skills, and abilities required to address items at a given level (Papageorgiou, Xi, Morgan, & So, 2015). Consequently, there is no unequivocal answer to the question of what it signifies to be at a particular level, as Harsch (2019) observes: "...classification of test-takers into proficiency levels necessitates human interpretation in addition to 'hard' statistical analyses" (p. 81). We therefore assert that the descriptors shaping the tests for validation clearly specify, in qualitative terms, the knowledge, skills, and abilities necessary to achieve a CEFR level. Although the 50% probability is not inherently characteristic of CEFR descriptors, we employed this metric with the aim of specifying expected values for the number of students at the CEFR level that the test is designed to assess. Recognizing its origins in Rasch scaling, we adopted the 50% probability to quantitatively evaluate learner performance using Chi-square analyses within the context of classical test theory. As indicated in Table 1, the probabilities for learners at a given level to successfully engage with tasks at a higher level range from .12 to .18; we therefore utilized this range in our statistical analyses as well.

In conclusion, by taking the probability assumptions and the argument-based validation approach as foundational elements, this study aimed to validate four-level tests of Turkish L2 receptive skills, guided by the following research questions:

1. Do the level tests of reading and listening classify students into intended CEFR levels?
2. Do the cut scores determined for each level test back the results of test scores?
3. Are the expected and observed number of students under and above the cut scores statistically significant to be accepted as warrants for the validity claim?
4. Are there any conditions that the abovementioned claims do not apply? If so, can they be supported with enough data to be considered as a counterclaim?

Method

This paper is part of a larger study aimed at developing descriptors of Turkish L2 proficiency for four skills at A1, A2, B1, and B2 levels. The comprehensive study unfolded in four phases: preparation, development, implementation, and validation. The present study, however, focuses on reporting the results from the implementation and validation phases and reveals several innovative methodologies in a descriptive manner.

Participants and Materials

The participants in the implementation phase were 384 students of Turkish L2 enrolled at the Turkish Language Center of Anadolu University. These university students were selected for two primary reasons: firstly, the researchers are affiliated with higher education and possess substantial experience working with students at this academic level; secondly, the CEFR descriptors are deemed especially suitable for young adults and adult learners (Benigno & De Jong, 2016). The students represented a diverse cohort from regions such as the Middle East, Eastern Europe, Africa, and Central Asia, and were pursuing undergraduate and graduate studies in Turkish universities. Many participants were either recipients of the Türkiye Scholarships program or self-funded. Excluding those from Turkic countries, most students lacked any instructional background in Turkish; therefore, they typically began at the A1 level based on a placement exam. The intensive program consisted of roughly 25 hours of language instruction per week throughout the year, divided into 7 or 8 weeks of modules for each level. To progress beyond each level, students must achieve a score of 70 or higher on comprehensive tests that include all four skills along with grammar and vocabulary; scores below 70 necessitate level repetition. After completing five levels (starting from A1), they are expected to reach the either B2 or C1 level by year's end to commence their respective university programs (Türkiye Bursları, n.d.). Several Turkish language centers offer an Academic Turkish course for students who successfully complete the C1 level, which correlates with the C2 level.

For the validation phase that included standard setting, language instructors (eight at the time of data collection) at the language center were involved. Half held undergraduate degrees while the other half had graduate degrees. Their teaching experience ranged from 9 to 23 years, equipping them with the practical insights necessary to judge learners' proficiency and the competencies needed to achieve a CEFR level in Turkish L2. They all participated in specification, familiarization, and standardization training as detailed in the CEFR Manual (CoE, 2009), outlined in the procedure section below.

As mentioned, this paper constitutes a segment of a larger project that addressed all four skill sets for Turkish L2 assessment. However, due to significant methodological differences between productive and receptive skills in the broader study—and our adoption of alternate approaches for validating productive skills—this study concentrates on receptive skills. Therefore, the materials comprised four level tests (A1, A2, B1, and B2) with listening and reading sections in a multiple-choice format, each evaluated out of a total of 100 points. The tests were developed based on the tables of specifications that aligned descriptors with test features, such as the number and length of texts and audio stimuli, item counts per task, and weighting. Assessment experts reviewed these specifications, appraising the extent to which test contents reflected the defined constructs. These findings were detailed in an earlier study (Savuran & Çubukçu, 2021). Piloting the tests on a separate group of learners from the same language center took place in prior academic terms, with preliminary item difficulty analysis prompting content edits to enhance certain items, particularly those with discrimination values below .10. The procedure section below offers additional details regarding the test implementations.

Sampling and Procedure

We administered level tests to mixed student groups, selecting participants through purposeful sampling. Each level test group comprised students from the preceding level, those of the target level, and those from the subsequent level—as previously explained in the 'Participants' section regarding student level determination. Table 2 presents the numbers of students from various levels who participated in the tests.

Table 2. Numbers of test-takers of four level tests

Test/Number of Students	Pre A1	A1	A2	B1	B2	C1
A1	25	66	23	-	-	-
A2	-	21	87	24	-	-
B1	-	-	25	88	27	-
B2	-	-	-	26	97	21

As Table 2 indicates, each level test was taken by three groups, the majority of whom were students of the intended level. For instance, the A2 level test was administered to 132 students, with 21 students from A1, 87 from A2, and 24 from B1 levels. There were two main reasons for including upper and lower levels in the sample. First, lower-level students were added so that the number of students falling under and above the cut score determined for that level could be observed. Second, upper-level students were included with the intention of balancing the score distribution and aiding in the item analysis of the test. Participation in each test was voluntary, and while obtaining their consent, we informed the students about the study and explained that the tests were for research purposes only, meaning that the scores would not be equated with or substituted for those of official level completion tests. However, the first author of the study, who was an instructor at the language center, encouraged the students to take the tests, stating that they would serve as real practice for their end-of-level tests. The implementation of the four tests lasted around two years under such conditions, with students who were willing to participate and assess their performance within their levels. The tests were administered in the 6th week of 8-week modules designated for the instruction of one level, with one exception for the A1 level test. As argued in relevant literature, determining the Pre-A1 level is challenging since it is regarded as a real-life proficiency level of a tourist nature in CEFR section 3.5. (CoE, 2001). However, the Pre-A1 level is interpreted as 'halfway to A1' (CoE, 2020, p. 243); therefore, students in the 3rd week of their A1 level instruction were considered Pre-A1 level students and were tested at different times than A1 and A2 level students.

To determine a cut score for each level test, the Angoff method was chosen since it is an item-centered standard-setting technique (CoE, 2009), and our aim was to evaluate the validity of the tests rather than individual examinee performance. Another reason for utilizing the Angoff method is that it allows users the flexibility to assign a cut score for each level test separately. Additionally, the Angoff method is one of the most commonly used standard-setting methods, not only for its simplicity and convenience in practice but also for the reliability of its results (Buckendahl, Smith, Impara, & Plake, 2002; Plake & Cizek, 2012). The standard-setting procedure was primarily facilitated by the first researcher and carried out virtually (due to COVID-19 restrictions) through video-conferencing and online forms for data collection. Based on Kane's three types of validity (internal, external, and procedural) (1994), steps outlined in the CEFR Manual (CoE, 2009) were followed diligently. Panelists reported no significant challenges with the virtual standard-setting process, except for a few technical issues. Moreover, collecting expert ratings online proved to be practical as it saved a great deal of time when exporting data to spreadsheets and conducting analyses, as suggested in relevant literature (e.g., Katz & Tannenbaum, 2014).

Since CEFR descriptors define the expected learner performance, the level definitions act as Performance Level Descriptors (PLDs) in a standard-setting meeting (Tannenbaum & Cho, 2014). For the familiarization stage, preparatory and introductory activities, together with a qualitative analysis of CEFR descriptors, were conducted to ensure a common understanding of CEFR level definitions, scales, and descriptors among the panelists. Tables of specifications, prepared during the test development process, were provided to the panelists. They were informed about the test preparation phase, e.g., how the assessment experts confirmed the alignment of the descriptors with the test content and the pilot testing process. Additionally, panelists were reminded that they should consider the difficulty of the tasks across levels. For example, identifying anaphoric relations in a text begins at the A2 level with basic words in simple and familiar texts; however, it is measured in extended texts at the B2 level. Emphasizing such transitions from simple to complex tasks across levels to the panelists was essential for the study since they were asked to consider the concept of a 'borderline examinee' (Cizek & Bunch, 2007) when setting the cut scores for each test. After discussing what constitutes a 'minimally competent candidate' at each level, we directed the panelists to the tables of specification and asked them to evaluate how well three types of test-takers (poor, average, and good) would perform the specified tasks across the levels so that they could better envision the 'borderline examinee.' A sample standard-setting practice for standardization training was conducted before the actual standard-setting for each of the four tests, which were done in two rounds. The panelists were asked, "Imagine 100 borderline examinees answered the item; how many of them would get it correct?" (CoE, 2009, p. 63). After the first round, the results were shared with the panelists to allow them to compare their own probability estimates with those of others. Following informal discussions, they went through a second round of judgments, the results of which were accepted.

Data Analyses

Although item response theory (IRT) is utilized in many recent test validation studies (McNamara & Knoch, 2012), the current study adheres to classical test theory (CTT) for data analyses for two main reasons. First, CTT assumes that observed scores are a combination of true scores and error, and it regards the standard error of measurement (SEM) as group-dependent, while IRT considers it as independent of the group (Magno, 2009). Second, IRT is considered more appropriate for larger sample sizes (Fulcher & Davidson, 2013). Since each test was administered to three different groups of learners, and the sample size is relatively small, we adhered to CTT in comparing score variances of groups by levels, as well as in the item analyses.

In order to provide solid backing for our validity claims, we first analyzed each test based on the normality of score distribution by calculating skewness, kurtosis values, and conducting Shapiro-Wilk tests on the overall scores. Subsequently, we analyzed the difficulty (p) and discrimination (d) indices of each item in the tests. The ideal p value was considered to be around .63, since items in our tests offer four options (Thompson & Levitov, 1985). For the item discrimination index (d value), items were categorized as 'very good' if their d value was above .40; 'good' if it ranged from .30 to .39; 'average' if it was between .11 and .29; 'poor' if it was .10 or below (Hopkins, 1998, p. 260). In addition to item-based difficulty and discrimination values, we calculated the overall item difficulty and discrimination for each test through the means of the p and d values of individual items in a test.

To determine a cut score for each test, panelists' probability estimates collected online were exported to spreadsheets. Their responses for each item in each of the two rounds were recorded below each other as shown in Table 3.

Table 3. Data Analyses Sample for Cut Score Determination

	Panelist									
Item / Round	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Med.**	**Std.**
1 / 1	.2	.25	.4	.54	.12	.27	.3	.35	.3	.13
1 / 2	.4	.38	.42	.52	.45	.32	.42	.45	.4	.09
2 / 1	.24	.28	.38	.41	.32	.58	.42	.75	.4	.17
2 / 2	.35	.45	.52	.4	.47	.45	.48	.39	.4	.06
Sum/ **Median**	17.1	16.7	18.7	19.1	15.6	16.5	14.7	21.4	**16.7**	**2.09**

Med: Median, Std: Standard Deviation

Table 3 displays a sample data analysis sheet used for determining the cut scores. The responses of the panelists in the second round were totaled and displayed in the bottom row for each panelist to review their individual cut score for that test (Cizek & Bunch, 2007). To consolidate these individual cut scores into a single common score, we chose to use the median, as Cizek and Bunch have recommended for standard-setting endeavors that employ the concept of the 'borderline examinee' (2007). The median of the eight individual cut scores was accepted as the proposed cut score for each test and was then converted to a score out of 100, as the tests were scored on this scale as well. It should be noted that these scores were kept fractional and not rounded, since our objective was to compare them with the test-taker scores, and not to use them for pass/fail decisions.

To provide validation for our claims, we employed the cut scores for each test to compare the test scores of students at two adjacent levels that the cut score is presumed to separate. For example, the proposed cut score for the B1 level test was utilized to compare the mean test scores of A2 and B1 level students taking it, in light of the observed and the expected numbers based on the probability values presented in Table 1. With this approach, we conducted the chi-square goodness of fit test, which analyzes whether there is a statistically significant difference between the observed distribution and the theoretical expected distribution (Frey, 2018). The degrees of freedom were set at 1, and significance was established at p < 0.05. Residuals, which indicate the discrepancy between the expected and observed number of students at two adjacent levels both below and above a cut score, were treated as rebuttal data and recognized as instances of false positives and false negatives. The internal validity of the standard-setting procedure was confirmed through Cronbach alpha (for the reliability of the panelists' judgments) and intra-class correlation (ICC) measurements (for the agreement consistency among panelists), as suggested in the CEFR Manual (CoE, 2009).

Results

We present the results of the three types of data we have; score distributions and psychometric properties of the tests, cut scores and their related reliability and ICC measures, and the results of chi-square goodness of fit tests.

Test Scores and Psychometric Properties

Before examining the scores of the three groups of test-takers for each level test, we conducted normality tests on the overall scores. The skewness values for each of the four tests, from A1 to B2, were -0.23, -0.22, -0.13, and -0.1, respectively, while the kurtosis values were -1.21, -0.89, -0.91, and -0.98, respectively. Significance (p) values for the Shapiro-Wilk tests at the level (α) of 0.05 were 0.000, 0.003, 0.004, and 0.002, respectively. The negative skewness and kurtosis values for all four tests indicate that the distributions are left-skewed and lighter-tailed. These p-values, being smaller than the 0.05 significance level for the Shapiro-Wilk tests, further confirm that the score distributions are not normal. This non-normality allowed us to infer that the scores of the three groups of test-takers in each test varied at statistically significant levels.

Following these statistical analyses, we aimed to observe how the scores of the three groups of test-takers differed from each other. Table 4 presents the number of students at three different levels who took the same exam, along with their mean scores and standard deviations.

Table 4. Score Distributions by Level

A1 Level Test				
	Pre-A1	**A1**	**A2**	**Overall**
No. of Students	25	66	23	114
Score Mean	38.8	64.65	80.1	62.1
Std.	11.83	18.78	10.32	21.06
A2 Level Test				
	A1	**A2**	**B1**	**Overall**
No. of Students	21	87	24	132
Score Means	45.23	62.95	79.89	63.21
Std.	9.93	19.07	13,15	19.67
B1 Level Test				
	A2	**B1**	**B2**	**Overall**
No. of Students	25	88	27	140
Score Means	41.15	60.71	74.25	59.69
Std.	13.08	21.9	16.79	22.05
B2 Level Test				
	B1	**B2**	**C1**	**Overall**
No. of Students	26	97	21	144
Score Means	42.07	63.4	83.09	62.67
Std.	11.2	19.15	10.39	20.41

Std: Standard Deviation

Table 4 presents the distribution of students' scores on the tests (out of 100) and overall score normality. A key finding is the consistent increase in mean scores by level for a specific test, which is evident across all four tests. For example, the mean scores of A1, A2, and B1 level students taking the A2 level test are 45.23, 62.95, and 79.89, respectively. The total number of students across the three levels taking that exam is 132, with an overall mean score of 63.21, almost identical to the A2 level students' mean score (62.95). The clear escalation of the mean scores by level in all four level tests suggests that the tests effectively reflect the intended learner performance.

The standard deviation (std.) of the scores for three levels of students further supports this claim. To illustrate, the overall std. for the B1 level test is 22.05, while it is 13.08 for A2, 21.9 for B1, and 16.79 for B2 level students. A higher std. value for students at the targeted level can be seen as evidence that the test measures the constructs of the test-takers in a more reliable way by differentiating between lower- and higher-performing students within that level (B1 level learners for the B1 level test). This finding is reinforced by the lower std. values for scores of students for whom the test is not intended (A2 and B2 students in the case of the B1 level test), indicating a more uniform distribution, which may suggest that the test is less relevant for them (i.e., the B1 level test not being as appropriate for A2 and B2 level learners).

Table 5. Test Reliability and Item Analysis

	KR-20	Item Difficulty (p) Interval	Overall Item Difficulty (p)	Item Discrimination (d) Interval	Overall Item Discrimination (d)
A1	.79	.35 - .78	.58	.11 - .58	.26
A2	.90	.31 - .82	.56	.11 - .51	.25
B1	.84	.33 - .86	.55	.12 - .53	.27
B2	.78	.16 - .84	.49	.16 - .55	.30

Apart from the score distribution and normality, we applied KR-20 analyses for reliability, as well as item difficulty (p) and discrimination (d) indices for item analysis on the four level tests. Although the KR-20 values for the A1 and B2 level tests are slightly lower than those for the A2 and B1 level tests, the reliability of all tests was found to be acceptable (Frey, 2018). The mean values of the item analysis based on each item's p and d indices can be seen in Table 5. The values of p for the items in the B2 level test, a rather different from the others, ranged between .16 and .86. The range of p for the A1, A2, and B1 level tests was between .31 and .86. However, the overall p values for the four tests were .58, .58, .55, and .49, respectively, indicating that, on average, about half of the examinees answered the items correctly in all tests. The d values for the items in the four tests ranged between .11 and .58. Nevertheless, the overall mean d values for each test were .26, .27, .25, and .30, respectively, which suggests that the tests' discriminatory effect between good and poor performing students was within the 'average' range, as stated in the relevant literature. This interpretation suggests that the tests are well-suited to differentiate test-takers based on their receptive skills within their respective levels.

The score distributions by level and the item analyses of the tests provide significant results regarding the tests' potential to reflect the intended learner performance. Score distributions, through skewness, kurtosis, and Shapiro-Wilk analyses—coupled with acceptable reliability and item analysis values across all four tests taken by three different levels of students—indicate that the tests were effective in measuring the learners' listening and reading abilities. It can thus be concluded that utilizing tests within the framework of the Assessment Use Argument (AUA) provided initial insights into their effectiveness. Through these solid psychometric properties, we were able to consider the test scores as 'data' (as specified in Figure 1 above), upon which we base our validity claim. To provide further backing from the data to support the claim, we implemented the Angoff standard-setting procedures for each test.

Standard Setting Results in Comparison with the Student Scores

After going through the familiarization, specification, and standardization training stages, which included sample standard-setting sessions, we asked eight panelists to provide their probability judgments for each item in a test. The analysis of their estimates in the second round was performed as illustrated in Table 3. The proposed cut scores for each test are thus presented in Table 6, along with the Cronbach alpha and ICC values for two rounds, identified as R1 and R2. The left side of Table 6 shows the scores of the students taking a level test (out of 100), while the right side displays the cut score, Cronbach alpha, and ICC values for that specific test.

Table 6. Standard Setting Results in Comparison with Test Results

	Test Results						**Standard Setting Results**				
	Student Score Means						**Cut**	**Cronbach alpha**		**ICC**	
	Pre-A1	**A1**	**A2**	**B1**	**B2**	**C1**	**Score**	**R1**	**R2**	**R1**	**R2**
A1	38.8	64.65	80.1				41.97	.71	.77	.64	.72
A2		45.23	62.95	79.89			50.6	.79	.91	.74	.83
B1			41.15	60.71	74.25		48.4	.83	.94	.76	.86
B2				42.07	63.4	83.09	46.7	.75	.89	.71	.84

Before comparing the mean scores of three groups of test-takers for each test with the cut score determined for that test, we examined the internal validity of the standard-setting processes. Cronbach's alpha and ICC values for each round of each test were found to be within the acceptable range (above .70), with an observed increase in the second rounds, indicating that panelists' judgments became more reliable and consistent. On the other hand, when looking at the students' scores from an overall perspective, it is clearly evident that the proposed cut scores, determined by the panelists through Angoff ratings, fall within the interval between the scores of students whom the test is intended to appeal to and those of the lower-level students whom the cut score is supposed to differentiate. For instance, the cut score determined through panelist judgments for the B1 level test is 48.4 out of 100, with the actual scores of A2, B1, and B2 level students for the same test being 41.15, 60.71, and 74.25, respectively. These findings allow us to interpret that the tests effectively measured the abilities of the students they were designed to assess. Additionally, the cut scores based on item-based judgments of the panelists support the test results (e.g., the cut score for the B1 level test [48.4] falls between the scores of A2 level [41.15] and B1 level [60.71] students). Therefore, we can conclude that cut scores can be viewed as evidence that 'support' the data (test scores) for the validity claim (the tests' ability to classify learners into the intended CEFR levels).

Another notable finding in the comparison of cut scores with student scores is that the cut score for the A1 level test is considerably lower than those for A2, B1, and B2 levels, which aligns with the student scores. The A1 level is theoretically and practically the foundation level in the CEFR, making it a logical anticipation for the panelists to assess the performance of a minimally competent candidate (a hypothetical borderline examinee) as the lowest among all four levels. Corroborating this pattern, the Pre-A1 level students, who took the test in their third week, scored the lowest of all 12 student score categories (the scores of three different level students for each of the four level tests). This may be because they were not yet familiar with the Turkish language in the third week of instruction, and thus they lacked the competency to perform well on the test.

Results of Goodness of Fit Tests Based on Cut Scores

The analysis of the mean scores of students taking a level test, in comparison with the cut score of that test, provided us with a clear picture of the impact of the tests on assessing the performance of the examinees. However, we sought to conduct further statistical analyses to scrutinize the number of students at adjacent levels who scored below and above the cut score, which is presumed to distinguish between them. To this end, chi-square goodness of fit tests for each level were run, based on the actual number of students taking the test and the expected number of students determined by the 50% probability assumption outlined in Table 1. The degrees of freedom (df) were set at 1, and a significance level (p) of <0.05 was used as the threshold for determining significance. The results of these analyses for the A1, A2, B1, and B2 tests are presented in Table 7.

Table 7. Chi-Square Goodness of Fit Tests for Level Tests

	Pre-A1			**A1**		
	Observed Number	**Expected Number**	**Residual**	**Observed Number**	**Expected Number**	**Residual**
Below Cut Score	14	22.0	-8.0	12	33.0	-21.0
Above Cut Score	11	3.0	8.0	54	33.0	21.0
Total		25			66	
Chi-Square		24.242			26.727	
Significance		.000			.000	
	A1			**A2**		
	Observed Number	**Expected Number**	**Residual**	**Observed Number**	**Expected Number**	**Residual**
Below Cut Score	15	17.2	-2.2	20	43.5	-23.5
Above Cut Score	6	3.8	2.2	67	43.5	23.5
Total		21			87	
Chi-Square		40.614			25.391	
Significance		.000			.000	
	A2			**B1**		
	Observed Number	**Expected Number**	**Residual**	**Observed Number**	**Expected Number**	**Residual**
Below Cut Score	17	22	-5.0	25	44.0	-19.0
Above Cut Score	8	3	5.0	63	44.0	19.0
Total		25			88	
Chi-Square		12.593			16.409	
Significance		.000			.000	
	B1			**B2**		
	Observed Number	**Expected Number**	**Residual**	**Observed Number**	**Expected Number**	**Residual**
Below Cut Score	16	22.0	-6.0	20	48.5	-28.5
Above Cut Score	9	3.0	6.0	77	48.5	28.5
Total		26			97	
Chi-Square		13.636			33.495	
Significance		.000			.000	

Table 7 illustrates the significant differences between the scores of learners at adjacent levels on level tests based on goodness of fit tests in a detailed manner. To delve deeper into the analysis, let us take the B1 level test as an example. The cut score for this test was determined to be 48.4. In total, 113 students—25 from the A2 level and 88 from the B1 level—taking the B1 level test were included for the chi-square analysis. Based on the probability values given in Table 1—12% probability for A2 level students to solve tasks at the upper level—we expected 3 out of 25 A2 level students to score above 48.4, leaving the remaining 22 to score below it. For B1 level students with a 50% probability of solving tasks at their own level, we expected 44 out of 88 to score above 48.4. However, the observed numbers for A2 level students scoring below and above the cut score were 17 and 8, respectively, resulting in a residual value of ± 5.0. For B1 level students, with expected numbers of 44 below and above the cut score, the observed were 25 below and 63 above, yielding residual values of ± 19.0. The results of the goodness of fit tests for A2 and B1 levels were 12.59 and 16.4 (df: 1), respectively, and the difference between the observed and expected numbers of students was statistically significant ($p < 0.05$) for both levels. We interpret that the cut scores determined by the panelists based on the test items significantly categorized the students into the correct levels according to their actual test scores. This leads us to conclude that chi-square analyses based on cut scores can serve as warrants for the validity claim—that the tests successfully classify learners into intended CEFR levels.

To sum up, we employed the Assessment Use Argument (AUA) and standard setting to validate the listening and reading skill tests for Turkish as a second language within an argument-based validity framework. The findings presented in this section support the validity claim that the tests indicate the proposed learner performance and accurately classify learners into the intended CEFR levels, based on the data of learner scores at three different levels for one level test. Cut scores determined through the Angoff standard setting process for each test provide 'backing' for the validity claim. Similarly, the statistical significance of the differences between the expected and observed numbers of students at each level act as 'warrants'. On the other hand, the residuals may be interpreted as false positives and false negatives, which can be considered as 'rebuttals' with concrete numbers serving as 'rebuttal data'. We present the possible reasons and explanations concerning the data and the validity claim, alongside the warrants and rebuttals in the discussion and conclusion section that follows.

Discussion

Based on Toulmin's seminal work (1958, 2003) on argument structure and Bachman and Palmer's (2010) Assessment Use Argument (AUA), Papageorgiou and Tannenbaum (2016) proposed four validity claims with corresponding warrants and rebuttals. Although they assert that validity claims are inferential and interconnected, we opted to focus on Claims 3 and 4, which regard interpretations of the assessed abilities and the assessment records. Since our objective was not decision-making based on test results, we did not evaluate Claims 1 and 2, which pertain to test consequences and decisions. As highlighted in the literature, researchers have the discretion to select the types of claims they wish to propose concerning test scores and to gather evidence supporting their claims (Im et al., 2019; Kane, 2013). Hence, in relation to Claim 3, we maintain that interpretations of assessing listening and reading abilities of students are indicative of the intended learner performance and can be largely attributed to—and generalized within—the target language use (TLU) domain, Turkish L2 in our case, as the tests were designed and specified based on descriptors underpinning learner performance. In the context of Claim 4, we assert that students' scores across different levels and task types are consistent, as evidenced by the comparison of scores among students at three levels for the four tests.

We used test scores and their related interpretations since evidence drawn from learner performance is considered a prerequisite for AUA (Im et al., 2019; Kane, 2013), and many validation studies in language testing heavily rely on evaluating learner performance within an argument-based validation framework (Becker, 2018; Chapelle et al., 2010; Knoch & Chapelle, 2018; Mendoza & Knoch, 2018; O'Loughlin, 2011). However, the number of studies that embrace standard setting for argumentation (Cizek & Bunch, 2007; Lavery et al., 2020; Papageorgiou & Tannenbaum, 2016; Shin & Lidster, 2017) or describe standard setting procedures for argument-based validity (Kenyon, 2012; Kenyon & Römhild, 2013) has been relatively sparse. This scarcity leaves space for researchers to craft specific validity arguments within comparatively novel frameworks (Davies, 2012; Knoch & Chapelle, 2018). Benefiting from cut scores derived from Angoff standard setting procedures within the AUA framework, this study provides practical implications for validation endeavors by utilizing probabilistic assumptions of CEFR level placement.

We employ the assumption of a 50% probability of being at a CEFR level (De Jong & Benigno, 2017), based on Item Response Theory (IRT) scaling, to offer statistical evidence concerning the cut scores determined for four level tests. Although the topic has been subject to debate (Harsch, 2019; Harsch & Hartig, 2015; Hulstijn, 2007) with no definitive answers, recent research on assigning CEFR levels has revealed some quantitative insights (De Jong & Benigno, 2016, 2017). It should be noted that, while we used 50% probability for tasks at the learner's own level and 12-18% probabilities for tasks at an upper level to seek statistical significance, these figures should be viewed as providing numerical implications and not be regarded as conclusive. As Hulstijn states, "the notion of language proficiency presented in the CEFR rests on two, closely intertwined pillars: quantity and quality" (2007, p. 663), indicating that the probabilities may reflect the quantitative aspects of a learner's proficiency level in terms of knowledge. For assessment purposes aligned with the CEFR, researchers and test developers must define skills and abilities to meet specific needs and objectives, addressing the qualitative aspect of the adopted proficiency approach (Harsch, 2014; Papageorgiou et al., 2015), and they should provide solid documentation and finer scoring reports instead of broad band levels (Harsch, 2019). However, few studies have empirically reported the extent of their success in this area and the general issues affecting such endeavors in particular contexts (Brunfaut & Harding, 2020). With respect to Turkish L2 assessment, this paper represents a novel study with its rigorously designed tests, validated within the AUA approach through standard setting.

Numerous standard setting studies have been documented since the 1960s in the field of educational measurement (Harsch & Kanistra 2020), and it is recognized that there is no 'gold' or 'true' standard (Cizek, 1993; Kane, 1994); furthermore, without a universal framework like the CEFR, boundaries between adjacent levels or grades may lack meaning for those outside the standard setting process (Harsch & Hartig, 2015). However, having determined cut scores for four tests (A1 to B2) across six CEFR levels (Pre-A1 to C1) and adhered to guidelines outlined in the Manual (CoE, 2009), we claim that the cut scores have the potential to accurately classify learners based on the intended CEFR levels within language assessment, particularly with reference to Turkish L2. As conveyed in the literature, the study context is crucial when considering the risks of high numbers of false positives and negatives (Papageorgiou & Cho, 2014; Xi, 2007), and optimizing cut scores to minimize these risks is more critical than their precision (Cizek & Bunch, 2007; Eckes, 2017). Therefore, our aim was not to classify learners or decide on their levels, but rather to empirically compare the number of students scoring below and above the determined cut scores. Moreover, the chi-square analyses did not show statistically significant residuals, which refer to false positives and negatives for any of the four tests, allowing us to conclude that rebuttals did not pose counterclaims within our AUA framework.

Conclusion and Recommendations

The current study aimed to validate tests of Turkish as a second language (L2) within an argument-based validation approach through standard setting. It embraces the perspective that "it is the scores and uses of the test that are validated, not the test itself" (AERA, 2014; Kane, 1994). By interpreting and comparing test scores with cut scores, this research introduces an innovative method for validating receptive skills tests with validity claims grounded in data, backings, warranties, and rebuttals, supported by statistical evidence based on the probabilities of being at or solving tasks of a CEFR level.

The study has several limitations. First, it represents results from a relatively small sample size of test-takers. Moreover, the participants were motivated by the first author to take the tests as practice for real examinations. Second, the study relies solely on classical test theory in its analyses and does not incorporate generalizability theory or item response theory. Third, the participants are limited to students and panelists from a single language center. While this enhances familiarity and consistency, particularly for the concept of the 'borderline examinee' in standard setting, it opens avenues for future research to replicate findings by re-testing different student groups and determining new cut scores with other panels of experts or alternative standard-setting methods to strengthen external validity. Another constraint is the limitation of the tests to adult learners and to the A1-B2 level range, excluding the C1 and C2 levels.

Given this is a validation study report, it emphasizes the approach and methodology rather than the specific findings. The study notably presents a pioneering step in validating tests in the context of Turkish L2 assessments within the AUA framework. Accessible literature does not yield any comparable studies for Turkish L2 to set against our findings. Some comprehensive systematic review studies (Chapelle & Voss, 2021; Im et al., 2019; Lavery et al., 2020) highlight that most validation studies have focused on mainstream languages rather than less commonly taught languages such as Turkish.

In conclusion, this paper should be regarded as an initial effort. Future research on developing and validating descriptors and tests for Turkish L2 that target various learner ages and proficiency levels would contribute to the burgeoning relevant literature. We strongly encourage researchers in the field of Turkish L2 to conduct similar studies and document their methodologies, enhancing the impact on assessment practices in this field. This approach would enable the implementation of more reliable formative and summative assessment procedures, thereby addressing the needs of an increasingly large number of students..

References

American Educational Research Association. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, *2*(1), 1-34. doi:10.1207/s15434311laq0201_1

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Becker, A. (2018). Not to scale? An argument-based inquiry into the validity of an L2 writing rating scale. *Assessing Writing, 37*, 1-12. doi:10.1016/j.asw.2018.01.001

Benigno, V., & De Jong, J. (2016). The "global scale of English learning objectives for young learners": A CEFR-based inventory of descriptors. In M. Nikolov (Ed.), *Assessing young learners of English: Global and local perspectives* (pp. 43-64). New York: Springer. doi:10.1007/978-3-319-22422-0_3

Brunfaut, T., & Harding, L. (2020). International language proficiency standards in the local context: Interpreting the CEFR in standard setting for exam reform in Luxembourg. *Assessment in Education: Principles, Policy & Practice*, *27*(2), 215-231. doi:10.1080/0969594X.2019.1700213

Buckendahl, C. W., Smith, R. W., Impara, J. C., & Plake, B. S. (2002). A Comparison of Angoff and Bookmark Standard Setting Methods. *Journal of Educational Measurement*, *39*(3), 253-263. Retrieved from http://www.jstor.org/stable/1435081

Chapelle, C. A., & Voss, E. (2014). Evaluation of language tests through validation research. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1079-1097). New York: Wiley. doi:10.1002/9781118411360.wbcla110

Chapelle, C. A., & Voss, E. (Eds.). (2021). *Validity argument in language testing: Case studies of validation research*. Cambridge: Cambridge University Press.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (Eds.). (2008). *Building a validity argument for the test of English as a foreign language*. New York: Routledge.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, *29*(1), 3-13. doi:10.1111/j.1745-3992.2009.00165.x

Cheng, L., & Sun, Y. (2015). Interpreting the impact of the Ontario Secondary School Literacy Test on second language students within an argument-based validation framework. *Language Assessment Quarterly*, *12*(1), 50-66. doi:10.1080/15434303.2014.981334

Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement, 30*(2), 93-106. doi:10.1111/j.1745-3984.1993.tb01068.x

Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods, 17*(1), 31-43. doi:10.1037/a0026975

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting*. Thousand Oaks, CA: Sage. doi:10.4135/9781412985918

Council of Europe. (2001). *Common European framework of references for languages.* Retrieved from https://rm.coe.int/1680459f97

Council of Europe. (2009). *Relating language examinations to the common European framework of reference for languages: Learning, teaching, assessment (CEFR)*. Retrieved from https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680667a2d

Council of Europe. (2020). *Common European framework of references for languages - companion volume.* Retrieved from https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4

Council of Higher Education. (2023). *Statistics*. Retrieved from https://istatistik.yok.gov.tr/

Council of Higher Education. (2024). *Scholarships for international students*. Retrieved from https://www.studyinturkiye.gov.tr/StudyinTurkey/ShowDetail?rID=KlqzJ6l8YDQ=&&cId=PE4Nr0mMoY4=

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621-694). Washington: American Council on Education.

Davies, A. (2012). Kane, validity and soundness. *Language Testing, 29*(1), 37-42. doi:10.1177/0265532211417213

De Jong, J., & Benigno, V. (2016). The CEFR in higher education: Developing descriptors of academic English. Paper presented at Language Testing Forum 2016 - University of Reading. Retrieved from https://ukalta.org/wp-content/uploads/2016/10/DeJongBenigno_LTF2016.pdf

De Jong, J., & Benigno, V. (2017). Alignment of the global scale of English to other scales: The concordance between PTE Academic, IELTS, and TOEFL. Pearson: Global Scale of English Research Series. Retrieved from https://www.pearson.com/content/dam/one-dot-com/one-dot-com/english/TeacherResources/GSE/GSE-Alignment-other-scales.pdf

Douglas, D. (2010). *Understanding language testing* (1st ed.). New York: Routledge. doi:10.4324/9780203776339

Eckes, T. (2017). Setting cut scores on an EFL placement test using the prototype group method: A receiver operating characteristic (ROC) analysis. *Language Testing, 34*(3), 383-411. doi:10.1177/0265532216672703

European Commission. (n.d.). *Erasmus+ EU programme for education, training, youth and sport.* Retrieved from https://erasmus-plus.ec.europa.eu/

Frey, B. (Ed.). (2018). *The SAGE encyclopedia of educational research, measurement, and evaluation*. Thousand Oaks, CA: Sage. doi:10.4135/9781506326139

Fulcher, G. (2015). *Re-examining language testing: A philosophical and social inquiry*. New York: Routledge. doi:10.4324/9781315695518

Fulcher, G., & Davidson, F. (Eds.). (2013). *The Routledge handbook of language testing*. New York: Routledge.

Gomez, P. G., Noah, A., Schedl, M., Wright, C., & Yolkut, A. (2007). Proficiency descriptors based on a scale-anchoring study of the new TOEFL iBT reading test. *Language Testing, 24*(3), 417-444. doi:10.1177/0265532207077209

Harsch, C. (2014). General language proficiency revisited: Current and future issues. *Language Assessment Quarterly*, *11*(2), 152-169. doi:10.1080/15434303.2014.902059

Harsch, C. (2019). What it means to be at a CEFR level. Or why my Mojito is not your Mojito - on the significance of sharing Mojito recipes. In A. Huhta, G. Erickson, & N. Figueras (Eds.), *Developments in language education: A memorial volume in honour of Sauli Takala* (pp. 76-93). Jyväskylä: University of Jyväskylä Centre for Applied Language Studies. Retrieved from https://www.ealta.eu.org/documents/resources/Developments%20in%20Language%20Education%20A%20Memorial%20Volume%20in%20Honour%20of%20Sauli%20Takala.pdf

Harsch, C., & Hartig, J. (2015). What are we aligning tests to when we report test alignment to the CEFR?. *Language Assessment Quarterly*, *12*(4), 333-362. doi:10.1080/15434303.2015.1092545

Harsch, C., & Kanistra, V. P. (2020). Using an innovative standard-setting approach to align integrated and independent writing tasks to the CEFR. *Language Assessment Quarterly*, *17*(3), 262-281. doi:10.1080/15434303.2020.1754828

Hopkins, K. D. (1998). *Educational and psychological measurement and evaluation* (8th ed.). Boston: Allyn & Bacon.

Hulstijn, J. H. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal, 91*(4), 663-667. Retrieved from http://www.jstor.org/stable/4626094

Im, G. H., Shin, D., & Cheng, L. (2019). Critical review of validation models and practices in language testing: Their limitations and future directions for validation research. *Language Testing in Asia, 9*. doi:10.1186/s40468-019-0089-4

Kane, M. T. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, *64*(3), 425-461. doi:10.3102/00346543064003425

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Washington: American Council on Education.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1-73. doi:10.1111/jedm.12000

Katz, I. R., & Tannenbaum, R. J. (2014). Comparison of web-based and face-to-face standard setting using the Angoff method. *Journal of Applied Testing Technology*, *15*(1), 1-17.

Kenyon, D. M. (2012). Using Bachman's assessment use argument as a tool in conceptualizing the issues surrounding linking ACTFL and CEFR. In E. Tschirner (Ed.), *Aligning frameworks of reference in language testing: The ACTFL proficiency guidelines and the Common European Framework of Reference for Languages* (pp. 23-34). Almanya: Stauffenburg Verlag.

Kenyon, D. M., & Römhild, A. (2013). Standard setting in language testing. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 944-961). Hoboken, NJ: John Wiley & Sons.

Knoch, U., & Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Language Testing*, *35*, 477-499. doi:10.1177/0265532217710049

Lavery, M. R., Bostic, J. D., Kruse, L., Krupa, E. E., & Carney, M. B. (2020). Argumentation surrounding argument-based validation: A systematic review of validation methodology in peer-reviewed articles. *Educational Measurement: Issues and Practice*, *39*(4), 116-130. doi:10.1111/emip.12378

Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *The International Journal of Educational and Psychological Assessment, 1*(1), 1-11. Retrieved from https://files.eric.ed.gov/fulltext/ED506058.pdf

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing, 29*(4), 555-576. doi:10.1177/0265532211430367

Mendoza, A., & Knoch, U. (2018). Examining the validity of an analytic rating scale for a Spanish test for academic purposes using the argument-based approach to validation. *Assessing Writing, 35*, 41-55. doi:10.1016/j. asw.2017.12.003

Messick, S. (1989). Validity. R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York, NY: American Council on education and Macmillan.

North, B. (2000). *The development of a common framework scale of language proficiency.* New York: Peter Lang.

O'Loughlin, K. (2011). The interpretation and use of proficiency test scores in university selection: How valid and ethical are they?. *Language Assessment Quarterly, 8*(2), 146-160. doi:10.1080/15434303.2011.564698

Papageorgiou, S., & Cho, Y. (2014). An investigation of the use of TOEFL® JuniorTM Standard scores for ESL placement decisions in secondary education. *Language Testing, 31*(2), 223-239. doi:10.1177/0265532213499750

Papageorgiou, S., & Tannenbaum, R. J. (2016). Situating standard setting within argument-based validity. *Language Assessment Quarterly*, *13*(2), 109-123. doi:10.1080/15434303.2016.1149857

Papageorgiou, S., Xi, X., Morgan, R., & So, Y. (2015). Developing and validating band levels and descriptors for reporting overall examinee performance. *Language Assessment Quarterly*, *12*(2), 153-177. doi:10.1080/15434303.2015.1008480

Plake, B. S., & Cizek, G. J. (2012). Variations on a theme: The Modified Angoff, Extended Angoff, and Yes/No standard setting methods. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 181-199). New York: Routledge.

Savuran, Y., & Çubukçu, Z. (2021). Yabancı dil olarak Türkçe öğretiminde performans betimleyicileri geliştirme: Temel ve ara düzeyler. *Türk Eğitim Bilimleri Dergisi, 19*(2), 831-856. doi:10.37217/tebd.876422

Shin, S.-Y., & Lidster, R. (2017). Evaluating different standard-setting methods in an ESL placement testing context. *Language Testing*, 34(3), 357-381. doi:10.1177/0265532216646605

Tannenbaum, R. J., & Cho, Y. (2014). Critical factors to consider in evaluating standard-setting studies to map language test scores to frameworks of language proficiency. *Language Assessment Quarterly*, *11*(3), 233-249. doi:10.1080/15434303.2013.869815

Thompson, B., & Levitov, J. E. (1985). Using microcomputers to score and evaluate items. *Collegiate Microcomputer*, *3*(2), 163-168.

Toulmin, S. (1958). *The uses of argument.* Cambridge: Cambridge University Press.

Toulmin, S. (2003). *The uses of argument* (Updated ed.). Cambridge: Cambridge University Press.

Türkiye Bursları. (n.d.). Hakkımızda. Retrieved from https://www.turkiyeburslari.gov.tr/about

Xi, X. (2007). Validating TOEFL® iBT Speaking and setting score requirements for ITA screening. *Language Assessment Quarterly*, *4*, 318-351. doi:10.1080/15434300701462796

Zwick, R., Senturk, D., Wang, J., & Loomis, S. C. (2001). An investigation of alternative methods for item mapping in the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice*, *20*(2), 15-25.