Assessing Reading in Turkish as a Second Language: Scoring and Criterion-Related Validity

Yavuz Kurt and Gülcan Erçetin

Abstract

Justifying the appropriateness of test use for a particular purpose is critical in language testing. The tools used for assessment need to be continually evaluated before and after their use. Arguments regarding the validity of interpretations based on test performances can be developed from various aspects. Two of these aspects are justification for scoring and comparison against external measures. The current paper reports on an investigation of scoring validity and criterion related validity with regard to a set of reading tasks developed for second language (L2) learners of Turkish. The findings provide a preliminary base to develop validity arguments regarding the tasks, but also call for revisions.

Keywords: Scoring validity, criterion-related validity, assessing reading in Turkish as a second language

Introduction

Validity is an essential trait that makes a particular assessment *useful* for the purposes it is intended to serve (Bachman & Palmer, 1996). Messick (1989a) defines validity as "the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores" (p. 13). Sources of such evidence are multiple, such as theory, content, or criterion related, but the concept has long been perceived as *unitary*; therefore, different sorts of evidence are accepted as complementary (Messick, 1989a, 1989b; Weir, 2005). Since the current views of validity hinge on the legitimacy of score interpretations, judgements regarding validity need to focus on the appropriacy of meanings attached to particular test performances, rather than the testing tool itself (Akbari, 2012; Weir, 2005).

Assessment tools are meant to measure a construct or a group of constructs defined within a theory (Alderson, 2000). Widely acknowledged frameworks concerning the assessment of reading center on reading purpose and cognitive processes involved in reading (Enright et al., 2000; Weir & Khalifa, 2008). However, most of the studies that aimed to investigate reading ability for assessment purposes are in English. Reading ability in Turkish as a second language (TSL) has scarcely been investigated as a theoretical construct, and even less so, from an assessment perspective. Following Weir's (2005) test validation framework, the current paper presents a posteriori evidence regarding the criterion-related validity and scoring validity of a set of reading tasks that aimed to assess reading ability in TSL.

Yavuz Kurt, Boğaziçi University, Department of Foreign Language Education, yavuz.kurt@boun.edu.tr
Gülcan Erçetin, Boğaziçi University, Department of Foreign Language Education, gulcaner@boun.edu.tr

Review of Literature

The Central Role of Validity in Assessment

"Removing as much uncertainty as possible" is how Fulcher and Davidson (2007) describe validation, a process of minimizing ambiguity so as to be more confident when making decisions based on test performances (p. 4). Such a process is not straightforward and requires building arguments concerning the relationship between test performances, their interpretation and use (Bachman, 2005). When sociocultural values that influence how people understand constructs and the *social consequences* of using tests are taken into account to have a more comprehensive understanding of validity (Messick, 1989a), it becomes evident that there are many ways to look for evidence regarding the validity of a test use since the concept is a multifaceted one (Weir, 2005).

Assessment is a prediction regarding the status of an individual in relation to a construct - an abstraction of what a given skill or trait means (McNamara & Roever, 2006). Such prediction requires making inferences about a test-taker's ability since it is not possible to have direct information regarding what an individual can do in the real-world environment. Therefore, the predictions we make should be justifiable (McNamara & Roever, 2006). Given that language tests have been a common mechanism in making important decisions for various purposes such as employment and admission to educational institutions, evidence in support of the validity of a test use is critical; otherwise, in the absence or inadequacy of such evidence, the assessment practices could lead to misinterpretation of results or unfair treatment towards test-takers (Bachman, 2004). The potential sources of validity evidence such as content representativeness, theoretical background, relationship with other measures and consequences of using an assessment tool, all contribute to score meaning and interpretation in the holistic view of validity (Messick, 1989a). Within this view, any kind of justification for test use is seen as part of gathering collective evidence; therefore, more evidence from more sources means less doubt to cast on score meanings.

The Construct of Reading

Reading is a complex construct, and it has been examined from various perspectives such as the processing direction it involves (e.g., bottom-up, top-down) or component subskills required in reading (e.g. decoding and language comprehension) (Gough & Tunmer, 1986). In the literature, the definitions of reading as a construct usually involves a "perception" and an "interpretation" process (Grabe & Stoller, 2002; Urquhart & Weir, 1998), and emphasize the reader factor to imply that the meaning construction is not just the function of the text (Goodman, 2001; Kintsch & Rawson, 2005; Koda, 2005). Moreover, the construct becomes more complex when we take into account the multiple factors that have been found to be related to second language (L2) reading ability. These factors include L2 proficiency (Bossers, 1991; Jiang, 2011), background knowledge (Leeser, 2007; Sabatin, 2013), vocabulary knowledge (Hsueh-chao & Nation, 2000; Qian, 2002; Zhang, 2012), metacognitive knowledge (McNeil,

2011), socio-economic status (Kieffer, 2010), and knowledge regarding rhetorical organization of texts (Carrell, 1987; Rozimela, 2014; Zhang, 2008).

Khalifa and Weir (2009) proposed a model of reading that has been influential in designing tests of reading. The model is an updated version of the original model proposed by Urquhart and Weir (1998) and puts cognitive processes involved in reading at its center along with the potential interaction of these processes with reading purpose, and linguistic and background resources (Ünaldı, 2010). Reading purpose determines the kind of reading strategy to be employed by the reader and the cognitive processes required to complete a reading task, for example a task might simply require gathering propositional meaning or inferencing based on a single text while another one might entail creating an intertextual representation of multiple texts (Khalifa & Weir, 2009; Weir & Khalifa, 2008). Obviously, different levels of linguistic, background and rhetorical knowledge are needed depending on the sort of text and task that the reader is engaged with (Weir & Khalifa, 2008). The model hypothesizes that reading is carried out either carefully or expeditiously, and at either global or local level depending on the purpose of reading (Khalifa & Weir, 2009). With expeditious reading, the aim is quickly read a part of a text in order to find a specific piece of information or to scan for specific items, or read the whole text in order to have a general understanding. With careful reading on the other hand, the reader aims to pay attention to detail and understand the whole content within a *micro-propositional* level (local) or *macro-propositional* level (global) (Khalifa & Weir, 2009).

The model also refers to a set of contextual parameters regarding task setting and linguistic demands that can potentially interact with reading purpose, cognitive processes and reader knowledge (Khalifa & Weir, 2009), as empirically shown in relation to, for example, the response format employed in tasks, order of items, channel of presentation, rhetorical mode of the text (Jaliehvand & Moses, 2014; Zhou, 2011), grammatical and lexical complexity (Hsueh-chao & Nation, 2000; Morvay, 2012) and topic of the text (Lee, 2007).

Scoring Validity

Scoring soundness of test performances is thought to be one of the sources of validity evidence. "Scoring validity" as suggested by Weir (2005) is preferred instead of the traditional term "reliability" in order to draw attention to the fact that the consistency of scores across divisions of a measurement is an inherent part of validity. Therefore, instead of conceptualizing reliability as a separate aspect of test use, it needs to be understood as a source of evidence under the broader concept of validity (Weir, 2005). Consistent measurement of an ability is a precondition to any claim regarding the validity of a test use (Alderson, Clapham, & Wall, 1995). The variance in scores that is not a result of variance in the trait being measured is accepted as measurement error; therefore, decreased measurement error is associated with higher reliability (Bachman, 2004). The data from test administrations inform us about how much trust we can put in scores in that sense (Weir, 2005). The scope of scoring validity is summarized by Weir (2005) as follows:

> Scoring validity concerns the extent to which test results are stable *over time, consistent in terms of the content sampling and free from bias*

> [emphasis in original]. In other words, it accounts for the degree to which examination marks are free from errors of measurement and therefore the extent to which they can be depended on for making decisions about the candidate. (p. 23)

Since investigating scoring validity of a test use depends on scores from a particular testing event, the interpretations should make reference to those particular tools employed and the test-taker sample (Weir, 2005). Weir (2005), in his test validation framework, presents various potential sources of evidence regarding each aspect of validity, and these sources change depending on the language skill in question. For example, inter-rater consistency becomes a viable option to investigate the scoring validity, only when the test task requires some sort of language production to be scored by multiple raters.

The framework proposed by Weir (2005) involves sources of evidence for context-related, theory-related, scoring-related, consequential and criterion-related validity, all of which are brought together in a way that represents the temporal and conceptual relationships among these sources. For reading tests, the framework suggests that scoring aspect of validity has four potential sources of evidence: *item analyses, internal consistency, error of measurement, and marker reliability* (p. 44). Of these, item analysis and internal consistency are directly related to the measurement of receptive skills, whereas error of measurement and rater reliability are more of a concern for productive skills.

Classical item analysis can help us detect and treat certain weaknesses concerning test usefulness (Bachman, 2004). Items of appropriate difficulty for the target group of test-takers and items that can discriminate well between high performers and poor performers are considered essential for consistent and dependable scores (Weir, 2005). Items that are poor in terms of discrimination power or that have inappropriate difficulty levels cannot properly reflect the ability which is purported to be measured; therefore, cannot be relied upon. Weir (2005) also emphasizes that items that are supposed to measure the same underlying trait should correlate with each other, which would then indicate that a particular test is internally consistent in measuring the trait under focus.

Criterion-Related Validity

Another aspect of validity is criterion related and it is traditionally categorized as either concurrent or predictive validity. The extent to which test performances correlate with a criterion measure contribute to the meaningfulness of the scores (Messick, 1995), and appropriacy of meanings attached to scores is associated with the usefulness of the information generated by a particular test (Weir, 2005). Within this scope, comparison between the tool under investigation and an external measure that is supposed to tap into the same construct of interest can provide evidence to justify the use of a test for a specific purpose (Weir, 2005). The criterion that is used to compare with the tool under focus needs to be a reliable indicator of the same trait. Alderson et al. (1995) state that the criterion could be an alternative test that has established itself to measure the construct in question, examinee's self-evaluation of language skills, or ratings by the examinees' teachers on the corresponding skills being measured (p. 177). In a similar

fashion, Weir (2005) suggests that criterion-related validity could be investigated by exploring the relationship between test performances and a well-established external measure, teacher evaluations of students or self-evaluations of students. He explains that since teachers have the chance to periodically observe their students in the classroom environment, they are usually well-informed about their language abilities. Therefore, depending on how precise a teacher's judgments are, the empirical relationship between a set of scores and teacher evaluations could bear valuable information about validity (Weir, 2005). As such, comparison of a measurement against external criteria that are justifiable in terms of their similarity and relevance is useful when investigating the validity of test-based assessment of language skills including reading ability.

Assessment of reading ability in TSL is a scarcely studied area. Therefore, L2 teachers of Turkish have very limited resources of language assessment, and the few examinations of TSL in the market do not make clear arguments in order to justify their use. Thus, empirical research in the assessment of TSL area is needed. The current paper reports preliminary evidence regarding scoring validity and criterion-related validity in relation to a set of reading tasks developed for users of TSL. To this end, the following research questions were investigated:

1. To what extent do the psychometric characteristics of the reading tasks provide evidence for scoring validity?
2. Do the data from test-taker performance on the reading tasks provide criterion-related evidence as compared with self-ratings of test-takers and evaluations by their teacher?

Methodology

Reading Tasks

The reading tasks in the current study were developed for international students who study in Turkey at higher education institutions and use Turkish for both educational and personal purposes. Therefore, in the process of preparing the reading tasks, the potential language needs of the target population were taken into account. The content and type of texts are those that students can encounter both in their educational environment and in their daily life. The topics were carefully chosen to avoid bias against race, gender or religious affiliation.

A total of five reading tasks were developed in the study (see Kurt, 2015 for the tasks) and these tasks aimed at four proficiency levels, specifically B1, B2, C1 and C2 as defined by Common European Framework of Reference (CEFR, Council of Europe, 2001). The texts were checked for syntactic complexity in the form of words per sentence, and for lexical complexity in the form of word frequency based on Turkish National Corpus Word Frequency Lists (Aksan, Aksan, Mersinli, Demirhan, & Yılmazer, 2012). The percentage of words beyond the most frequent 2000 words was calculated for each text (see Table 1).

Table 1. Reading texts

Text	Intended level	Field	Text type	Word count	Words per sentence	Percentage of words out of the first 2000 list
1	B1	Environment	Expository	386	12.22	21.9%
2	B2	Cinema	Expository	333	15,04	32.2%
3	C1	History	Expository	419	14,31	33.2%
4	C2	Literature	Narrative	451	13,54	35.6%
5	C2	Biology	Descriptive	181	12,14	35.8%

The design of the tasks consisted of a process that involved i) the evaluation of perceived needs of learners; ii) identifying relevant reading strategies and skills based on the reading model by Khalifa and Weir (2009); iii) comparing the skills to CEFR can-do statements and forming specifications through sampling relevant reading skills; iv) selection of appropriate texts; v) item writing and item trial under the supervision of experts. The reading skills in Turkish that are required in a university environment could be diverse. For example, students might need to grasp an educational text for academic purposes or out of personal interest, but they might also want to have access to certain recreational materials such as literary works or magazines. Therefore, in accordance with the purpose of the test, such needs were taken into account in the planning phase of the tasks (Downing, 2006). A test based on such tasks could serve to assess the reading abilities of individuals in higher education in Turkey. The reading model (Khalifa & Weir, 2009) defines the scope of reading skills domain depending on the various reading purposes. For example, in an attempt to find specific information, test-takers are expected to scan a text and when they are asked to make inferences, they are expected to read a part or parts of the text. The test tasks were aligned to CEFR scales based on the assumption that certain reading skills required by task items can be expected only at or above a certain proficiency level defined by CEFR. To illustrate, based on the B2 level descriptive sentence "Can obtain information, ideas and opinions from highly specialized sources within his/her field, and can read articles and reports concerned with contemporary problems in which the writers adopt particular attitudes or viewpoints" it was assumed that test-takers at this level and above can distinguish fact from opinion and can understand author's attitude and viewpoint. The tasks reported here have 6 to 11 items each that essentially aimed to activate careful reading at local and global levels. While the first four tasks have multiple choice item format, information transfer is required in the fifth task where test-takers were asked to label the parts of a visual depending on the text they read.

Participants

The tasks were completed by a total of 62 international students at an English-medium state university in Turkey. The students were in Turkey at the time of the study through a student exchange program. On average, their length of stay in Turkey prior to the study was 14 months (SD = 31.5 with a range between 0-204 months). The participants were provided with a self-evaluation form on which they could evaluate their reading ability in Turkish on a six-level scale (see appendix). These levels correspond to six

proficiency levels according to CEFR (i.e. A1, A2, B1, B2, C1 and C2) and they involve a summary of Can-do descriptions so that students can easily position themselves on the form. As CEFR defines i) A1 level mostly comprises everyday familiar words and sentences; ii) A2 users can read short texts and locate predictable information in everyday materials; iii) B1 level is where users can independently deal with familiar texts regarding work, school or leisure; iv) B2 level is where users can understand viewpoints and attitudes, as well as technical discussions in their field; v) C1 level users are considered proficient and can read complex factual and literary texts or specialized articles in or out of their field; and vi) C2 level users can understand nearly all types of written text. At the time of the study, all participant students were enrolled in Turkish for Foreigners courses offered by the Turkish Language and Literature Department of the university. The students were taking these courses at one of two different proficiency levels. The course instructors reported that one group consisted of advanced level learners while the other one consisted of intermediate level learners. The students were placed in these courses depending on individual interviews conducted by the department instructors in order to evaluate students' proficiency. The course instructors also informed that students were sometimes moved to the higher or lower level proficiency group depending on their performance in the initial classes in the first few weeks.

Procedures

Each task was separately completed by 31 students in the classroom environment within the allocated time. Not all tasks were completed by all 62 participants since the tasks were delivered in two sets (Set A: Task 1, 2 and 4, Set B: Task 3 and 5) because of practical concerns. Durations allotted for each task were previously determined depending on feedback from a smaller sample of students. Before students took the tasks, they were asked to evaluate their reading ability in Turkish on the self-evaluation form. The data from student responses to the task items and the self-evaluation form were submitted to SPSS for statistical analyses. For the first research question that aimed to investigate scoring validity, two sources of evidence were investigated, i.e. classical item analysis and internal consistency. For item analyses, difficulty and discrimination indices were calculated for each item. Internal consistency was examined through calculating Cronbach's alpha values for each task. Suggested criteria for the item difficulty values usually do not go beyond the .20 - .80 interval (Bachman, 2004) while a minimum discrimination value of .20 is considered necessary to keep an item in use or for improvement (Frisbie & Ebel, 1991).

For the second research question that aimed to investigate the relationship between the tasks and external measures, the performances of test-takers were compared against two different criteria. The first one was based on the evaluation of students' instructors when assigning them to the higher level or lower level proficiency groups. Therefore, the performance of students on the tasks was investigated to reveal whether the tasks could effectively discriminate between the higher proficiency group and the lower proficiency group in alignment with the teachers' observations. The second criterion was students' self-ratings of their reading proficiency in Turkish on the CEFR

scale. The relationship between these ratings and their performance on the tasks were investigated.

Results

Given that the participant students are distributed along a proficiency continuum from intermediate to advanced levels, it was predicted that certain tasks could be below or above the proficiency level of test-takers, affecting the distribution of indices under investigation. Table 2 provides descriptive information regarding score characteristics. The mean scores on Table 2 indicate the average scores of students on each task, which should be evaluated taking into account the maximum possible scores since the score range for each task is different. On the other hand, mean difficulty indices are standardized, thus comparable across tasks. Values closer to 0 mean that students generally found the task difficult, whereas values closer to 1 mean the opposite. Therefore, although the highest mean score belongs to Task 2, the mean difficulty indices show that the students performed best on Task 5.

Table 2. Summary of score characteristics

Task	Intended level	Mean score	Maximum possible score	SD	Mean difficulty index
1	B1	2.19	6	1.92	.36
2	B2	4.93	11	3.57	.45
3	C1	2.96	8	2.13	.37
4	C2	3.19	8	2.19	.25
5	C2	2.03	7	2.16	.46

Note. N=31.

Table 3 presents the psychometric characteristics of items on each task. The column labeled as item-total correlation represents the discriminatory power of items, and the column left to it shows how the internal consistency would fluctuate assuming that the item was deleted. Most of the items across the tasks were within favorable difficulty levels. However, certain items on Task 1, 4 and 5 have difficulty indices lower than the suggested criterion .20, which means that less than 20% of the test-takers were able to provide the correct answer on these items. Therefore, these items seem to be too challenging for this group of test-takers. Additionally, there were no items that the test-takers found too easy to answer. Regarding item-total correlations, the majority of items effectively discriminated between high and low achievers. Nevertheless, a few items on Task 2, 3 and 4 were found to have item-total correlation values below .20, suggesting that these items performed poorly in terms of detecting poor and successful test-takers. Finally, the Cronbach's alpha levels indicate how consistent the items were within each task. All of the tasks had alpha values closer to .80 except Task 3. Cronbach's alpha levels if item deleted indicate how the internal consistency of the task would change if the item was removed from the task. Therefore, if this value is lower for a specific item when compared to the other items on the task, such as Item 1 on Task 4, this means the item positively contributes to the internal consistency of the task.

Table 3. Item statistics for reading tasks

Task number	Intended level	Item	Difficulty	Item-total correlation	Cronbach's alpha if item deleted	Cronbach's alpha for the task
1	B1	1	.48	.41	.770	.774
		2	.16	.57	.733	
		3	.58	.49	.749	
		4	.23	.58	.726	
		5	.39	.46	.755	
		6	.35	.64	.707	
2	B2	1	.45	.57	.860	.871
		2	.26	.63	.856	
		3	.45	.59	.858	
		4	.32	.55	.861	
		5	.35	.44	.868	
		6	.58	.60	.857	
		7	.68	.71	.851	
		8	.55	.69	.852	
		9	.42	.76	.846	
		10	.45	.61	.857	
		11	.42	.18	.886	
3	C1	1	.45	.41	.646	.685
		2	.26	.36	.659	
		3	.48	.66	.581	
		4	.32	.19	.696	
		5	.29	.25	.683	
		6	.58	.17	.704	
		7	.29	.50	.627	
		8	.29	.50	.627	
4	C2	1	.29	.64	.727	.779
		2	.35	.53	.746	
		3	.39	.24	.799	
		4	.10	.17	.794	
		5	.19	.61	.735	
		6	.19	.61	.735	
		7	.32	.48	.757	
		8	.19	.61	.735	
5	C2	1	.58	.57	.736	.778
		2	.58	.53	.745	
		3	.26	.57	.738	
		4	.65	.46	.758	
		5	.19	.51	.751	
		6	.52	.49	.753	
		7	.42	.41	.769	

To examine the criterion-related validity of the test, an independent samples t-test was run to compare the performances on the reading tasks of the two proficiency groups to which the students were assigned based on their teachers' evaluation (see Table 4).

Table 4. Comparison of the two proficiency groups

	Lower proficiency		Higher proficiency			
	M	SD	M	SD	t	Cohen's d
Task 1	1.33	1.72	3.00	1.79	2.64*	.95
Task 2	2.87	3.04	6.88	2.94	-3.73**	1.34
Task 3	2.00	2.04	3.76	1.92	-2.47*	.89
Task 4	0.71	1.07	3.12	2.26	-3.65**	1.36
Task 5	2.53	2.36	3.81	1.9	-1.67	

Note. * $p < .05$, ** $p < .001$, Lower proficiency (Set A) N=15, Lower proficiency (Set B) N=14, Higher proficiency (Set A) N=16, Higher proficiency (Set B) N=17.

The results showed that the first four reading tasks were efficient in discriminating the two proficiency levels since the performance of the higher proficiency group was significantly better than the other group. On the fifth reading task, although the higher proficiency group obtained higher scores on average than the lower proficiency group, this difference was not significant.

Another measure for the second research question was the relationship between test-takers' self-ratings of their proficiency and their performance on the tasks. Table 5 and 6 present the Pearson-product correlation coefficients between self-ratings of students (M=3.38, SD=1.19) and their scores on the tasks. As self-ratings are on a 1 - 6 scale representing A1 to C2 levels on the CEFR, the mean 3.38 indicates that students rated themselves approximately mid-way between B1 and B2 on average. Since half of the students completed tasks 1, 2 and 4 (Set A) while the other half completed tasks 3 and 5 (Set B), correlations between these two groups of tasks are not available.

Table 5. Correlations between self-ratings and task scores for Set A

Measure	Self-rating	Task 1	Task 2	Task 4
Self-rating	1			
Task 1	.589**	1		
Task 2	.718**	.700**	1	
Task 4	.382*	.449*	.621**	1

Note. N=31, * $p < .05$, ** $p < .001$.

Table 6. Correlations between self-ratings and task scores for Set B

Measure	Self-rating	Task 3	Task 5
Self-rating	1		
Task 3	.553**	1	
Task 5	.737**	.547**	1

Note. N=31, $^{**}p < .001$.

The correlation between self-ratings and task performances were found to be significant and positive across all tasks, with small to medium effect sizes. This indicates that the performance of test-takers on the tasks tends to reflect their self-rated level of proficiency.

Discussion

This study investigated scoring validity based on two relevant sources of evidence suggested in Weir's (2005) validation framework. Specifically, item analyses and internal consistency were examined based on data from test-taker performances on the reading tasks. Such analyses are useful in terms of revealing how consistent the obtained scores are (Bachman, 2004; Weir, 2005), which is associated with construct validity as a unitary concept (Messick, 1995). The results revealed that the majority of items were within acceptable difficulty range and had effective discriminatory power. A few items on certain tasks were found to be too challenging or weak in terms of discrimination for this group of test-takers. Such unfavorable difficulty and discrimination values can be a function of students' level of proficiency as well as *construct irrelevant* factors (Messick, 1989a), and thereby these could be considred measurement error (Bachman, 2004). In order to improve the reading tasks, these items were either replaced with new ones or revised through certain processes such as simplification of wording or employing less challenging distractors. Additionally, Item 4 and 6 on Task 3 had low discriminatory power with a negative impact on the internal consistency of the task. These items involved options that were intended to be challenging through processes such as the use of similar content words from the text or the use of peripheral propositions suggested in the text. This probably misled the test-takers to the incorrect options. It seems that endeavors to develop strong distractors might lead to malfunctioning distractors, causing an unfavorable situation where higher achievement group getting distracted to the incorrect options more frequently than the lower achieving group. These two items and a few others where a distractor was found to attract as many responses as the correct option draw attention to the importance of the quality of the options in multiple choice item format (Haladyna, Downing, & Rodriguez, 2002; Hughes, 2003).

Interestingly, Task 1 (B1 level), on average, was found more difficult than four of the other tasks intended for higher levels, which might be due to the the contextual features of the task such as text length, lexical and grammatical complexity, rhetorical features as well as the topic, all of which are associated with how cognitively

demanding a reading text is (Khalifa & Weir, 2009). Therefore, the text was revised in terms of syntactic and lexical complexity to obtain a simpler one, which decreased the word count, words by sentence ratio, and the number of infrequent words. Similar textual revisions were also carried out on the second task as well depending on the feedback from experts. An additional finding that should be emphasized is that although Task 5 was favorable in terms of internal consistency, item-total correlations and acceptable difficulty values, it was, overall, found easier than the other tasks. This task incorporated a relatively short and technical text on biology with advanced lexical items. Lexical complexity is associated with how demanding a text is (Hsueh-chao & Nation, 2000; Morvay, 2012). However, it seems that the lexical complexity of this text might have been neutralized by other potential factors including text length, sentence length, rhetorical mode and response format. The text was shorter than the others although it was on a specialized topic. It is worth taking into account that texts with specialized technical topics tend to incorporate a straightforward language, i.e. shorter sentences with very limited tense change, as opposed to literary works (as in Task 4) which usually have sentences with more increased length and with more tense shifts. Therefore, the test-takers might have found the text syntactically less challenging compared to the other text that aimed the same level (Morvay, 2012). Besides, test-takers might have found the supportive visual helpful in interpreting the information in the text (Khalifa & Weir, 2009), thereby transferring the words from the text to label the parts of the visual with little effort, and reducing the task to local expeditious reading. It is also possible that the reason was the expository mode of the text; expository texts were found less challenging than narrative ones in previous studies (Zhou, 2011). Therefore, this could be the reason why test-takers generally performed better on Task 5 compared to Task 4. These findings imply that such features of texts should be taken into account in text selection process.

The Cronbach's alpha values for three of the tasks were found to be at the upper end of .70-.80 interval. While Task 2 was above this interval, Task 3 was a little below it. Usually, internal consistency values below .70 are considered unfavorable (Fulcher & Davidson, 2006); however, the fact that the number of items on a test influences the consistency of scores should be taken into consideration (Fulcher & Davidson, 2006; Hughes, 2003). Given that this initial administration was on a small group and the number of items on each task was limited, the obtained values can be considered reasonable, and accepted as preliminary evidence for internal consistency of the tasks. The values indicate that the items on each task tend to measure the same underlying trait. Upon the revisions on the tasks with concerns of text complexity, item difficulty and discrimination, they could potentially bear more favorable results. In short, this line of analyses provided information regarding the dependability of scoring. The obtained data were utilized to improve the tasks in an attempt to decrease measurement error (Bachman, 2004; Weir, 2005).

A second set of analyses examined the relationship between test-taker performances on the reading tasks and external measures to investigate criterion-related validity. Weir (2005) suggests that the validity of scores can be checked against an external measurement, along with other alternatives such as an alternative version of the test, the same test, or future performance of individuals. One of these external measures employed in this study was the proficiency group to which the participant students

belong. These groups were identified as advanced or intermediate levels, and students were assigned to them depending on teachers' evaluation. Results from four of the five tasks, i.e. except Task 5, indicated that the two proficiency groups were efficiently discriminated. The finding that the task scores reflected the external measure, i.e. the two proficiency levels, contributes to the trustworthiness of score interpretation (Messick, 1995). However, the findings also imply that Task 5 could be problematic and is in need of revision or replacement. Furthermore, scores on all tasks were found to significantly correlate with students' self-ratings of their proficiency, with mostly moderate strength. Although usually very high correlations are expected between the obtained scores and the criterion measure, the individual reading tasks in this study are indeed limited performance samples. Therefore, larger samples of reading performance, for example in the form of several reading tasks combined to form a complete reading test, will probably bear higher correlations with external criteria. Still, the significant performance differences between the two proficiency groups of test-takers, and the same groups' perceptions of their own reading ability in Turkish are in line with the reading ability as reflected in test taking performance, giving support to accurate operationalization of the skill in the tasks. Consequently, this second set of analyses signal potential problems on Task 5 and provide arguments from a criterion-related aspect to justify the use of the remaining four tasks as a measure of reading ability in L2 Turkish (Weir, 2005).

Conclusion

This study presented findings regarding the scoring validity and criterion-related evidence based on student performances on five individual reading tasks. This initial administration of the tasks bore preliminary evidence that justify the interpretations of test-takers' reading ability in Turkish. However, the data also signalled needs for revisions or replacements on certain occasions. These tasks were also examined in terms of content and distractors, which are reported elsewhere (see Kurt, 2015). Examining testing tools from multiple aspects of validity can provide valuable data that is useful to improve the tools. The overall findings point to the fact that the skills required by the tasks, contextual factors and the intended specific test-taker group interact with each other forming a complex phenomenon. It was also revealed that characteristics of reading texts and the task design require utmost care in test development process.

The results reported here should be treated as tentative findings since this study is limited in a number of ways including sample size and sampling of reading skills, text types and item format types. Future researchers involved in the assessment of L2 reading in Turkish and any other language should take into account the multiple factors involved in reading assessment, some of which are treated in this paper. Different item formats and reading task designs should be studied in order to sample more reading skills on various types of texts. Only with more studies, development of tools that best elicit the abilities we aim to measure will become possible.

Acknowledgement

This paper is based on the unpublished Master's thesis titled "Development of a reading test for second language learners of Turkish" by Kurt (2015). We deeply appreciate the helpful and cooperative attitudes of Ceyda Arslan and Bilgen Erdem who teach in Turkish Language and Literature Department of the university where the data for this study were collected.

References

Akbari, R. (2012). Validity in language testing. In C.Coombe, P.Davidson, B.O'Sullivan & S.Stoynoff (Eds.). *The Cambridge Guide to Second Language Assessment* (pp.30-36). Cambridge: Cambridge University Press.

Aksan, Y., Aksan, M., Mersinli, Ü., Demirhan, U. U., & Yılmazer, H. (2012). *Turkish national corpus (TNC) demo version work frequency lists* (Report No. 1). Mersin: Mersin University.

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

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

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

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly: An International Journal*, *2*(1), 1-34.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Bossers, B. (1991). On thresholds, ceilings, and short-circuits: The relation between L1 reading, L2 reading and L2 knowledge. In J. H. Julstijn, & J. F. Matters (Eds.), *AILA Review, 8*, (pp. 45-60). Amsterdam: Free University Press.

Carrell, P.L. (1987). Content and formal schemata in ESL reading. *TESOL Quarterly 21*, 461-481.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Downing, S. M. (2006). Twelve Steps for Effective Test Development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3-25). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Enright, M., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, P., & Schedl, M. (2000). *TOEFL 2000 reading framework*. Princeton, NJ: Educational Testing Service.

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

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*. London, England: Routledge.

Goodman, K. S. (2001). *On reading*. Portsmouth, NH: Heinemann.

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education*, *7*(1), 6-10.

Grabe, W., & Stoller, F. L. (2002). *Teaching and researching reading*. London: Longman.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-334.

Hsueh-chao, M. H., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*, 403-430.

Hughes, A. (2003). *Testing for language teachers*. New York, NY: Cambridge University Press.

Jaliehvand, M., & Moses, M. (2014). Influence of rhetorical pattern on improving EFL sudents' reading comprehension. *Journal of Studies in Social Sciences, 7*, 210-225.

Jiang, X. (2011). The role of first language literacy and second language proficiency in second language reading comprehension. *The Reading Matrix, 11*, 177-190.

Kieffer, M. J. (2010). Socioeconomic status, English proficiency, and late-emerging reading difficulties. *Educational Researcher*, *39*(6), 484-486.

Khalifa, H., & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading.* Studies in Language Testing 29. Cambridge: UCLES/Cambridge University Press.

Kintsch, W., & Rawson, K. A. (2005). Comprehension. In M. J. Snowling, & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 209-226). Malden, MA: Blackwell.

Koda, K. (2005). *Insights into second language reading*. New York, NY: Cambridge University Press.

Kurt, Y. (2015). *Development of a reading test for second language learners of Turkish* (Unpublished master's thesis). Boğaziçi University, İstanbul, Turkey.

Lee, S. K. (2007). Effects of textual enhancement and topic familiarity on Korean EFL students' reading comprehension and learning of passive form. *Language Learning*, *57*(1), 87-118.

Leeser, M. J. (2007). Learner-based factors in L2 reading comprehension and processing grammatical form: Topic familiarity and working memory. *Language Learning, 57*, 229-270.

McNamara, T. F., & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA: Blackwell.

McNeil, L. (2011). Investigating the contributions of background knowledge and reading comprehension strategies to L2 reading comprehension: an exploratory study. *Reading and Writing, 24*, 883-902.

Messick, S. (1989a). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

Messick, S. (1989b). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5-11.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749.

Morvay, G. (2012). The relationship between syntactic knowledge and reading comprehension in EFL learners. *Studies in Second Language Learning and Teaching, 2*, 415-438.

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: an assessment perspective. *Language Learning 52*(3), 513–536.

Rozimela, Y. (2014). The students' genre awareness and their reading comprehension of different text types. *International Journal of Asian Social Science, 4*, 460-469.

Sabatin, I. M. (2013). The effect of cultural background knowledge on learning English language. *International Journal of Science Culture and Sport, 1*(4), 22-32.

Urquhart, S., & Weir, C. J. (1998). *Reading in a second language: Process, product and practice*. New York: Longman.

Ünaldı, A. (2010) *Investigating reading for academic purposes: sentence, text, and multiple texts* (Unpublished doctoral dissertation). University of Bedforshire, Bedforshire, England.

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

Weir, C. J., & Khalifa, H. (2008). A cognitive processing approach towards defining reading comprehension, *Cambridge ESOL: Research Notes, 31*, 2-10.

Zhang, X. (2008). The effects of formal schema on reading comprehension: An experiment with Chinese EFL readers. *Computational Linguistics and Chinese Language Processing, 13*(2), 197-214.

Zhang, D. (2012). Vocabulary and grammatical knowledge in L2 reading comprehension: A structural equation modeling study. *The Modern Language Journal, 96*, 554-571.

Zhou, L. (2011). Effects of text types on advanced EFL learners' reading comprehension. *Journal of Language and Culture, 30*(2), 45–56.

Türkçenin İkinci Dil Olarak Değerlendirilmesi: Puanlama Geçerliği ve Ölçütsel Geçerlik

Özet

Dil sınavlarının belli amaçlar için kullanımının uygunluğunun gerekçelendirilmesi önemlidir. Ölçme değerlendirmede faydalanılan araçların, kullanımlarından önce ve sonra sürekli olarak değerlendirilmesi gerekir. Test performanslarına dayalı çıkarımların geçerliğine ilişkin argümanlar çeşitli yönlerden geliştirilebilir. Bunlardan ikisi puanlamanın temellendirilmesi ve dış ölçütlerle karşılaştırmadır. Bu makale, Türkçe'yi ikinci dil olarak öğrenenler için geliştirilen bir dizi okuma görevinin puanlama geçerliği ve ölçütsel geçerlik açısından incelenmesini sunmaktadır. Bulgular, görevlerle ilgili geçerlik argümanları geliştirmek için bir ön temel sağlamakta, aynı zamanda görevler üzerinde düzeltmeler yapılması gerektiğini göstermektedir.

Anahtar sözcükler: Puanlama geçerliği, ölçütsel geçerlik, ikinci dil olarak Türkçe'de okuma becerisinin değerlendirilmesi

Appendix

How would you rate your reading ability in Turkish in the following areas? Please put a tick on the relevant box.

	A1	**A2**	**B1**	**B2**	**C1**	**C2**
Reading	I can understand familiar names, words and very simple sentences. ☐	I can read very short, simple texts. I can find specific information in simple everyday material such as advertisements, prospectuses, menus and timetables and I can understand short simple personal letters. ☐	I can understand texts that consist mainly of high frequency everyday or job related language. I can understand the description of events, feelings and wishes in personal letters. ☐	I can read articles and reports concerned with contemporary complex problems. I can understand contemporary literary prose. ☐	I can understand long and complex factual and literary texts. I can understand specialized articles and longer technical instructions, even when they do not relate to my field. ☐	I can read with ease virtually all forms of the written language, including abstract, structurally or linguistically complex texts, factual and literary texts, such as manuals, articles and literary works. ☐