Striking the Balance between Validity and Reliability of a Listening Test in Turkish as a Second Language

Emel Tozlu and Aylin Ünaldı

Abstract

Evidence on the efficacy of an assessment tool is necessary in order to justify the decisions we make based on the scores from it. Validity evidence can be collected from several sources such as the stages before and after test administration. In the present research study, validity evidence of several types on a Turkish as a Second Language (TSL) Academic Listening Test is presented in order to establish the efficacy of it. This paper presents cognitive, contextual and scoring validity (reliability) evidence from the first and second versions of the test and investigates whether the modifications made after the first administration have had a positive effect on the quality of the test. The study concludes that although the changes made in the first version of the test strengthened the validity claims in terms of cognitive and contextual requirements, the reliability scores of the test worsened in the second version. This reminded us that although it is necessary to build the foundations of a test firmly by operationalizing the necessary contextual features and cognitive processes, this will not thoroughly guarantee the technical quality of the items. Scoring validity should be established carefully as well. This study exemplifies a thorough attempt in establishing the validity of a TSL test from multiple perspectives and aims to be an exemplary study for further test development in TSL.

Keywords: Cognitive validity, contextual validity, scoring validity, assessment of listening in Turkish as a second language.

Introduction

It has been established in the field of language assessment that tests should go through determined stages of quality assurance both before and after test use. From each of these stages, it should be possible to derive validity evidence to support the accuracy and adequacy of the interpretations we make based on test scores (Messick, 1989). Kane (2013) argues that validation is an evaluation of the coherence and completeness of the arguments concerning test interpretation and use; it is a process of gathering evidence in support of the plausibility of our inferences and assumptions based on test scores. According to Kane (2013) validity of a proposed test interpretation or use depends on ‘‘how well the evidence supports the claims being made’’ (p.1). In any test development and test use process, there is a need to provide justification for the decisions made at each step. This is how scores from that test can be accepted as accurate indicators of the construct in question.

Several guidelines make it clear that a test should go through rigorous processes of validation before the scores from it can be generalized to performances in the target situation (i.e., ALTE, 2011; EALTA, 2006; Young, So & Ockey, 2013). Weir (2005) delineates a clear language assessment validation framework on which validity arguments at a priori and a posteriori stages can be developed; Weir’s framework

Emel Tozlu, Boğaziçi University, School of Foreign Languages, emel.hakyemez@gmail.com
Aylin Ünaldı, Boğaziçi University, Department of Foreign Language Education, aunaldi@boun.edu.tr

provides guidance on the kind and depth of the evidence that can be gathered in different phases of test development and test use. The main components of the framework are *test taker characteristics*, *cognitive* (theory-based) validity, *context* validity and *scoring* validity. Cognitive validity is related with adequate and appropriate operationalization in tests of the mental processes, which are normally prompted by the authentic language tasks in real life. Context validity concerns the accurate representation of task parameters such as textual features and the control of externally determined conditions. Scoring validity is related with the accuracy and reliability of the evaluation process through which a test taker is assigned a score. According to Weir (2005), these three aspects form the overall *construct* validity of any test. The other two components of this framework, *criterion-related* and *consequential* validity, are seen as dependent on construct validity.

This article will present a part of a larger study in which a newly developed Turkish as a Second Language (TSL) Academic Listening Test has been validated from several aspects. The part that is reported here is limited to the construct validation of the test; it discusses evidence on cognitive, contextual and scoring facets. Special emphasis is placed on the intricate balance that has to be maintained between validity parameters and the statistical reliability of the test. Although validity and reliability are widely discussed as two important qualities of a good test, many tests with high reliability quotients are accepted as well-functioning tests without strong validity evidence. However, cognitive and contextual validity can be low when reliability is high, or when cognitive and contextual aspects of tasks are paid attention to, the reliability may not be ensured. There are hardly ever any studies that discuss the ways to ensure both the reliability and the cognitive and contextual validity of a test at the a priori and a posteriori stages, especially when TSL tests are concerned. This study will, therefore, exemplify such an attempt by focusing on the following specific questions:

1. What are the cognitive requirements of the listening test tasks?
    a. Is the listening construct operationalized in the test tasks in a way that targets a sufficient range of cognitive processes indicated by the listening frameworks across different proficiency levels as predicted by the CEFR?
    b. Do the test takers' perceptions of the listening sub-skills that they employed to answer the items support that the test tasks can tap into the necessary cognitive processes?
2. What are the contextual characteristics of the listening test tasks?
    1. What are the demands imposed upon the test takers by task setting, administration setting, linguistic features of the listening test tasks and the speakers?
    2. What are the participants' perceptions of the tasks in terms of the suitability of their contextual features for the different proficiency levels?
3. How well do the test tasks and the items function in terms of scoring validity?
    1. Do the values for central tendency measures of the tasks and item analyses based on the test takers' performances support that the test is functioning well?

2. Does the test measure the listening ability of learners of TSL reliably?

Literature Review

A validation study has to situate itself in a validation paradigm so as to ensure systematicity. In Messick's (1989) unified conception, validity is defined as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment" (p. 13). This view has laid the grounds for modern understanding of the concept of validity; however, not without criticism. Messick's unitary view was seen as too broad and ambitiously integrative without practically applicable help to testers and researchers (Borsboom, Mellenberg, & Heerden, 2004; Knoch & Elder, 2013). Walt and Steyn (2008) underline that the lack of clearly separable categories make Messick's framework hard to use. Weir's (2005) socio-cognitive validity framework adopts a similar approach to the conceptualization of validity as Messick's; it is not the test but the inferences made based on the test scores obtained from a particular administration of a test on a particular group of test takers are validated, and validation is an ongoing process. However, thanks to the practical and detailed guidance on each language skill, Weir's framework is more applicable by practitioners.

Weir's Socio-Cognitive Framework

Weir's (2005) socio-cognitive framework for validating language tests outlines details for four language skills (reading, writing, listening and speaking) with reference to internal cognitive processing, external contextual factors, and individual test taker characteristics. Weir (2005) maintains that evidence collection for validity claims should be done both before (a priori) and after (a posteriori) the administration of the test emphasizing that all the components of the socio-cognitive framework complement each other in gathering substantial validity evidence. However, within the scope of this study we will only investigate cognitive, context and scoring validity (reliability) types.

Cognitive Validity

Cognitive validity, previously named as "theory-based validity" (Khalifa & Weir, 2009) is concerned with whether a test requires a test taker to be engaged in the cognitive or mental processes that are reflective of those that a listener would normally employ in a real-world listening situation (Weir, 2005). This means that the cognitive skill(s) that would be needed in accomplishing the test task should accurately and adequately represent the cognitive skills required by authentic 'language use' tasks. Thus, direct representation of real-life cognitive skills in the test tasks is necessary to attain cognitive validity. While investigating cognitive validity, working with a detailed and explicit definition of the language construct is essential. Therefore, a theoretical framework that underlies the construct needs to be used as the basis of test development. Following this principle, the TSL listening test analyzed in this study has been based on Field's (2013) listening framework, which will be discussed in more detail below.

Context Validity

Context validity, or as commonly cited 'content validity', refers to how representative the test tasks are of the target language use domain to which we would like to make generalizations (Weir, 2005). Context validity is mainly concerned with the authenticity of test tasks, i.e. with the degree of similarity between the characteristics of the task input and the expected output and the characteristics of non-test language use. Bachman and Palmer (1996) state that there are two different kinds of authenticity: (1) situational authenticity and (2) interactional authenticity. Situational authenticity of a test task depends on the degree of correspondence between the test method characteristics and the features of the target language use domain. On the other hand, interactional authenticity is dependent on the interaction between test takers and the cognitive processes they need to employ in order to fulfill the task. Therefore, it can be stated that situational authenticity is concerned with context validity while interactional authenticity is related to cognitive validity. Weir (2005) argues that these two validity types should be seen as complementary to each other since the contextual features of the test tasks influence the cognitive processes involved during test administration. Therefore, these two should be discussed in relation to each other. In Weir's (2005) framework, context validity is investigated in terms of task setting, administration setting and task demands in terms of linguistic features of the tasks and interlocutors. These three main components of context validity have several sub-components and each of these should be taken care of in order to be able to collect evidence for context validity.

Scoring Validity

Weir (2005) considers reliability as another aspect of validity; therefore, he refers to it as 'scoring validity' in his socio-cognitive framework.[2] Weir (2005) defines reliability as "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). McNamara (2000) states that reliability explores how good the process of assessment is by investigating scores. Geranpayeh (2013) maintains that reliability is related to data quality. While validity is concerned with the accuracy of the interpretations of the scores obtained from a test, reliability is related to "the consistency of the measurement" (Bachman & Palmer, 1996). Bachman (1990) points out that validity and reliability can be seen as related to two complementary aims in test design and development: "(1) to minimize the effects of measurement error [reliability], and (2) to maximize the effects of the language abilities we want to measure [validity]" (p. 161).

The sources of inconsistency and error in language tests should be identified and their effects should be minimized in order to prevent them from impacting the test scores and inevitably their use and interpretation. McNamara (2000) argues that in order to ensure the meaningfulness and fairness of the assessment, some quality control

[2] In this study, we will use the terms "scoring validity" and "reliability" interchangeably.

procedures need to be followed, one of which is item analysis[3].

Field's (2013) Framework for Listening

As previously mentioned, the language construct under investigation should be defined clearly in the light of a theoretical framework in order to have a sound basis for the test. Field's (2013) framework for the listening skill is composed of five "levels of analysis" and explains the listening comprehension process thoroughly. These levels are divided as lower-level processes (i.e., input decoding, lexical search, parsing) and higher-level processes (i.e., meaning construction, discourse construction. Input decoding refers to the comprehension of individual sounds and the attempt of the listener to match them against the phonological system of the language. The formation of a representation of the sounds is followed by searches in the lexicon for the syntactic function and the meaning of the word. Then, the listener assigns syntactic structures to the parts of the incoming utterance. The listener attempts to form a meaningful proposition in his/her mind, which carries the literal meaning of the utterance. Once a literal proposition is formed, the listener uses his/her world knowledge, topic knowledge and contextual clues in order to understand the real, intended meaning of the utterance. In other words, at the meaning construction stage, the propositional meaning, which reflects the literal meaning of the utterance, is transformed into the actual meaning of the utterance with the help of the knowledge that is available to the listener. After the intended meaning of the utterance has been formed in the listener's mind, all the utterances that have been said so far become a part of the listener's memory by being combined, analyzed and synthesized and the listener forms a discourse representation. In Field's (2013) listening framework, the processes of listening comprehension are not necessarily considered to be sequential; rather, this process view of listening comprehension emphasize the integrative nature of listening by underlining the interaction between the lower and higher level processes.

CEFR Descriptors for the Listening Skill

In addition to the listening framework proposed by Field (2013), the descriptors for the listening skill specified in the Common European Framework of Reference for Languages[4] (Council of Europe, 2001) can be integrated into a test development process so as to be able to define the listening construct and determine the target language sub-skills to be assessed. CEFR provides a common basis for the development of syllabuses, course books, classroom materials and examinations thus facilitates the standardization of such products. The language proficiency levels in the CEFR are defined with respect to all language skills and sub-skills from A1 to C2 levels. In this study, these sub-skills indicated for each proficiency level are taken as the basis to determine the target sub-

[3] In this study, classical test theory will be followed because it is still a widely used technique and the number of participants is restricting the use of a more advanced technique.

[4] The CEFR is developed by the Council of Europe (2001) in an attempt to describe comprehensively what language learners need to learn in order to be able to communicate through a language and what knowledge and skills they need to develop to be able to do so.

skills to be operationalized in the test tasks.

In sum, the present study makes an attempt to provide arguments for the validity of the score interpretations from the TSL Academic Listening Test by situating it in Field's (2013) listening framework and the CEFR. It follows Weir's (2005) validation suggestions in gathering evidence systematically. We present detailed accounts of the two versions of the test, before and after substantial revisions done to strengthen its cognitive and contextual properties, and the resultant effect of them on the reliability of the test. We aim to illustrate the steps that can be taken to improve a listening test and at the same time in order to provide evidence supporting the cognitive, contextual and scoring validity of the TSL Academic Listening Test. We hypothesize that the revisions done on the first version of the test will have a positive effect on the test quality and therefore give us strong evidence for validity and the reliability of the test.

Methodology

The TSL Academic Listening Test under investigation was prepared by one of the researchers and piloted with a group of TSL learners. However, although the initial item statistics were favourable (see Tozlu, 2017), due to some perceived validity concerns regarding the test tasks and items, the test was revised together by both researchers and underwent substantial changes before the second piloting. The test specifications, tasks and items were examined in terms of the components of Weir's (2005) validation framework, which guided us in arguing for the cognitive, contextual and scoring validity claims of the TSL Academic Listening Test. In the analysis, the test tasks from both versions of the test were examined to justify the modifications made after the first piloting and to show whether the tasks in the second version meet the necessary cognitive requirements. In addition to the theoretical discussion of the cognitive aspect of the test tasks, the task evaluation questionnaires provided data for the cognitive skills used by the test takers to answer each item in the second version. Context validity claims of the test were explored theoretically using the criteria in Weir's (2005) socio-cognitive test validation framework. Each component in his framework was examined separately and the appropriacy of the test tasks, items and test specifications were evaluated accordingly. Only significant results from this examination were discussed in the current study for practical reasons.

The First Administration of the Test

In the first administration, the test was given to 55 Erasmus students who were taking the Turkish for Foreigners (TKF) classes offered at Boğaziçi University and were at the proficiency levels ranging from B1 to B2 levels. The data were gathered through the administration of the five listening test tasks, which were aimed at different proficiency levels from A1 to C1 and involved a range of item formats such as gap-filling, multiple-choice and short-answer questions. During the preparation of the test, a number of TSL course books, syllabi for TSL courses and listening tests for TSL learners were examined and finally the test specifications were created and the test tasks were written in accordance with the test specifications.

After the first piloting, measures of central tendency and dispersion were calculated and classical test theory analyses were conducted using IBM SPSS 21 Software. Mean, range, and standard deviation, Cronbach's alpha (α), item discrimination (Corrected Item-Total Correlation-CITC), and reliability estimates for individual test items (Alpha If Item Deleted; AIID) were calculated to evaluate the effectiveness of the individual items (See Tozlu, 2017).

After the first administration of the test, the test was evaluated theoretically scrutinising its cognitive and context validity. Also taking the findings from the statistical analysis into consideration, the necessary alterations were made accordingly on the first versions of the tasks. Cognitive validity claims were based on the examination of the listening skill demands of the test tasks and test specifications in accordance with the premises of Field's (2013) listening framework, Weir's (1993) listening taxonomy and the CEFR specifications for listening.

The Second Administration of the Test

In the second administration, the new test tasks were given to a different group of 30 Erasmus students at a university, who had a variety of proficiency levels ranging from A1 to B2+. However, not all of the test tasks could be administered to all of the students due to time limitations and restrictions required by the TSL instructors. Therefore, A1 and A2 level students were exempted from B2 level task and likewise, B2 and B2+ students were exempted from A1 level task. This led to two different groups of participants in the second administration; i.e. lower-level test takers (A1, A2 and B1 level students) and higher-level test takers (B2 and B2+ level students).

During the modifications after the first administration, C1 task was eliminated and only four test tasks at A1, A2, B1 and B2 levels were included in the second administration. In addition to the test tasks, task evaluation questionnaires (see Tozlu, 2017) were distributed to the test takers and these provided valuable data regarding the test takers' perceptions of the test tasks. The test takers were asked to fill in the relevant questionnaires right after each test task was administered. In these questionnaires, they were requested to choose the cognitive skills that they employed for each item during listening from a set of cognitive skills provided in the questionnaire, and to evaluate the difficulty levels and the contextual features of the test items. After the administration of the second version of the test, the same classical test theory statistical procedures were conducted as in the first administration. In addition, the data from the task evaluation questionnaires were analyzed and integrated into the discussion.

The theoretical discussions and results of the statistical analyses carried out for both administrations of the test enabled researchers to derive validity and reliability evidence for the TSL Academic Listening Test and how these two facets of test validation interact with each other.

Results and Discussion

Cognitive Validity

For the investigation of cognitive validity, each test task both from the first and second administration was evaluated in terms of the underlying cognitive skills it aims to operationalize under the light of Field (2013), Weir (1993), and the CEFR. Furthermore, the results of the first section of the task evaluation questionnaire are discussed for each task in the second administration.

A1 Level Task

A1 level task in the test aimed to assess one listening sub-skill, which was "to listen for specific factual information clearly stated". In A1 level task, the test takers were supposed to answer six open-ended questions with short answers such as numbers or one or two word phrases. This task aimed at processing specific word-level information, thus lexical search as the cognitive process. Based on this, it can be stated that A1 level task, which was designed as the easiest task in the test, assessed lower-level processing in Field's (2013) framework as well as Weir's (1993). The CEFR specifications indicate that at A1 level listeners are assumed to understand simple, high frequency vocabulary when uttered slowly and clearly. The answers for the A1 level task in this study are composed of simple and high frequency words and the listening text was recorded at a slow and understandable speed. Therefore, it can be concluded that this task was designed to measure only lower-level processes and understanding factual, simple and clear information.

However, although the task was considered to assess the target sub-skill, one aspect of the test task that underwent changes was the nature of the information the items required. In the first examination, there were four questions which needed numerical answers and two questions which demanded content words. In order to create a balance, one of the questions was replaced by a question that required comprehension of a content word. By doing this, we aimed to achieve a better construct representation, yet still targeting comprehension of clear lexical information.

Table 1 in the appendix shows the cognitive skills that the test takers thought they employed most while listening according to the task evaluation questionnaires. In Table 1, the most popular sub-skill marked by the test takers for all of the items is the first sub-skill "understand specific bits of information in the dialogue", which is what this task precisely aims to tap into. It can also be seen that some extra sub-skills such as the fourth and fifth sub-skills were also employed. The listeners might have felt the need to employ additional cognitive processes in order to reach an effective understanding of the listening text and to compensate for their lack of linguistic knowledge. The prominence of the fourth sub-skill, 'Differentiating between important and less important information' might suggest that there might be distractors in the text, and the use of the fifth skill, 'understand what the dialogue is about briefly' suggests that at least some items in the task might have required more than local level processing. As it will be discussed below, item statistics confirmed these observations.

A2 Level Task

Initially the task was designed as a B1 level task; however, the statistical analyses and the analysis of the cognitive requirements of the target level suggested that the task was much easier than expected and with some adjustments it could be used as an A2 level task. On the other hand, the A2 level task in the first administration was later developed into a B1 level task in the second piloting due to similar problems experienced with the B1 level task in the first piloting.

According to Field's (2013) and Weir's (1993) frameworks and the CEFR specifications, listeners, at this level, are able to understand specific details in the listening text, and thus, lexical and sentence-level factual information. The new A2 level task was designed to target these cognitive skills. In addition, it was also modified to target direct meaning comprehension section: "Listening for main idea(s) or important information: and distinguishing that from supporting detail, or examples". Thus, A2 level task was differentiated from A1 level task in terms of its cognitive load. However, after the analysis of the test items in the second administration, it was observed that almost all of the items target lexical information, not sentence-level information. Moreover, the first question seemed to require inferencing skill, which is also inappropriate for the target level. These mean that if the first question is modified to target only sentence-level factual information and some others are also modified to do so, the cognitive validity claims of the test task will be stronger.

Table 2 in the appendix shows that the lower-level group mainly employed the first and fourth sub-skills while answering the questions. The target sub-skill for this task was to listen for specific information at lexical level for items 2 to 8; thus, the items seemed to have elicited the necessary sub-skills since both the first and fourth sub-skills are related to comprehension of specific details. However, for the first item, the test takers also marked other sub-skills (the second, fifth and seventh sub-skills). This shows that the item failed to measure only sentence-level factual information and it might have required some inferencing skills on the information in the text and on the speaker's attitude and tone.

According to Table 2 in appendix, the higher-level test takers mostly employed sub-skills related to understanding specific information (the first and third sub-skills). As opposed to the lower-level group, the higher-level group did not need to differentiate between important and less important information, which was, indeed, not necessary to carry out the task. Instead, they focused on the details used to understand the main ideas. However, they similarly attempted to understand the topic of the text for the first item by employing the fifth sub-skill. These results demonstrate that the test takers generally adopted sub-skills relevant to understanding specific information. This supports the assumption that this task assesses specific lexical information.

B1 Level Task

The B1 level task in the first administration was changed and turned into an A2 level task in the second administration as explained above due to cognitive and statistical concerns. Similarly, analysis of the A2 level task in the first piloting demonstrated that some of the items targeted cognitive processes inappropriate for this level. For certain

items in the A2 level task in the first administration, especially the first, third and sixth items, some of the answers were not directly stated in the listening text and the test takers were required to make inferences using the information in the text, which is considered to be a higher-level process during meaning construction (Field, 2013). Therefore, since the task was beyond the expected difficulty level, it was turned into a B1 level task in the second piloting.

At B1 level, learners are expected to understand beyond clear, simple and factual information and assessing extended discussions which trigger higher level listening processes is crucial according to both Field's (2013) and Weir's (1993) listening frameworks as well as the CEFR specifications. However, the first B1 level task failed to achieve this as it only targeted assessing "listening for specific information". Therefore, the second B1 level task included radical changes to better assess the listening construct. The new B1 level task aimed to assess the sub-skills such as:

- Listening for specifics, including recall of important details (Weir, 1993)
- Listening for main idea(s) or important information and distinguishing that from supporting detail, or examples (Weir, 1993)
- Understanding discourse markers (Weir, 1993)
- Identifying and reconstructing topics and coherent structure from ongoing discourse involving two or more speakers (Richards, 1983)
- Determining a speaker's attitude or intention towards a listener or a topic (Weir, 1993)
- Making inferences and deductions at local levels (Weir, 1993)

Therefore, we can conclude that a B1 level task is differentiated from an A2 level task in terms of its cognitive difficulty and is also much more appropriate for the target level.

Table 3 in the appendix shows the results of the task evaluation questionnaires obtained from the lower and higher-level test takers. The lower-level test takers employed a wider range of sub-skills to respond to the items in B1 level task since the first, third, fourth and sixth sub-skills were utilized the most by the lower-level test takers as well as the eighth sub-skill for the first and fifth items. The higher-level test takers mostly utilized similar sub-skills (the first, third and fourth sub-skills) as the lower-level test takers except for the sixth sub-skill; however, the popularity of the sub-skills for each item is not as strong as for the lower-level group. The overall findings indicate that the test takers employed higher-level listening processes such as main idea construction and comparing important and less important information. Therefore, it can be argued that these results reflect that the items in this task achieve to measure higher-level listening processes.

B2 Level Task

The first B2 level task was not used in the second administration of the test and a completely new task was created by the researchers because of cognitive and item format-related concerns. The B2 level task in the first administration required listeners to listen to an announcement about a university course and answer some questions in True/False/Not Given format. One of the major problems was the cognitive skills that the task targeted. The task did not tap into listening skills such as inferences, meaning

construction and discourse representation, which are essential at higher levels of listening. In addition, it did not assess listening to complex and extended speech on both concrete and abstract topics, which is necessary at B2 level according to the CEFR specifications. Neither the test items nor the listening text allowed for the assessment of these crucial listening skills. Besides, the item format was prone to guessing, a construct-irrelevant factor. Furthermore, the "Not Given" response alternative was not a suitable question format for the listening skill as trying to find out the missing information could create a great difficulty for the listener. Owing to these problems, the task was replaced by a new one which aimed to tap into higher-level listening sub-skills also targeted by the B1 level task with the addition of inferencing at both local and global levels.

Table 4 in the appendix demonstrates that the test takers reported the use of a variety of sub-skills for this task. Most of the higher-level test takers heavily utilized the first, third, fourth and eighth sub-skills for almost all of the items and employed the fifth and ninth sub-skills for some of the items in the task. This finding provides evidence for the level of variation across tasks in terms of the cognitive processes required and supports the theoretical suggestions.

C1 Level Task

The C1 level task in the first piloting was neither revised nor included in the second administration of the test due to a number of reasons. The task aimed to assess the academic listening skills of TSL learners; thus, it included an authentic lecture on information technologies, specifically computers. However, the analysis of the listening text and the items demonstrated that the text was heavily based on factual information and did not contain extended speech forms of discussions, which was not suitable for the target proficiency level. In addition to this, the linguistic features of the text were not at the expected level of difficulty either. This showed that genre, on its own, is not an indicator of difficulty level and rhetorical purpose and organization of information should also be taken into consideration while choosing a listening text (Weir, 2005). Since the B2 level task designed for the second piloting was considered to be sufficient to assess the listening skill at a higher level, no new C1 level task was prepared.

To conclude the investigation of cognitive validity, it can be stated that the levels of variation in listening skills across different levels of proficiency are better displayed in the second version of the test thanks to the modifications made after the first administration. This became evident through the comparison of listening sub-skills operationalized in the test tasks with the ones suggested in theoretical models (Field, 2013; Weir, 1993) and a language framework, the CEFR. We combined this observation with the listening skills that the test takers reported in the questionnaires and item statistics. We consider this as the evidence for the successful operationalization of the listening sub-skills in the test.

Context Validity

After the first administration, the test tasks were revised by following Weir's (2005) parameters for context validity in his socio-cognitive validation framework. Whether the

tasks succeeded in complying with the requirements of context validity was explored by considering each parameter of context validity in the discussion below. A summary of the findings is presented and the significant results are explained in detail (See Tozlu, 2017 for a complete examination of the test tasks in terms of context validity parameters).

Task Setting

Task setting covers the aspects of authenticity, response format, known criteria, weighting, order of items and time constraints. In the first version of the test, only the C1 level task had a fully authentic recording, which was a lecture about information technologies. However, after its careful examination, it was observed that the listening text, despite being authentic, did not demonstrate satisfactory features in terms of linguistic complexity; therefore, the C1 level task was completely discarded from the test. In the second version of the test, no fully authentic texts were used; however, the listening script for the B2 level task was a slightly revised authentic radio interview and recorded again due to the modifications made by the researchers. The texts for the other tasks both in the first and second administration were written by the researchers. Fully authentic texts could not be used due to the difficulty of adapting them for tasks such as multiple-choice. The instructions given to the text takers are also a part of authenticity in context validity conception. In the first administration, the test takers were given the instructions in both written and spoken forms; however, the written instructions provided information only about the speakers and the completion of tasks whereas the spoken ones also informed the test takers about how many times they would listen to the recordings and how much time they were given to go over the questions before listening. In the second version, in addition to these, details regarding the setting and the topic were also added to both written and spoken instructions in order to create context for the listeners.

Response format considers task types, answer keys, linguistic difficulty of item stems, possibility of note-taking, and memory load. In the first administration, short answer questions were written for the A1 and B1 level tasks, multiple-choice questions for the A2 level task, "True/False/Not Given" questions for the B2 level task and fill-in-the-blanks questions for the C1 level task. In the new version, two main task types, short-answer questions for A1 and A2 levels and multiple-choice questions for B1 and B2 levels were utilized. For short-answer questions, issues such as alternatively correct answers, acceptable spelling mistakes and length of answers were considered more carefully and detailed answer keys were prepared. Only a limited range of spelling mistakes were accepted as long as they did not yield a new meaningful word and mostly short, high frequency words were targeted as the correct answer, which complied with the requirements of the CEFR specifications at these levels. For multiple-choice questions, items with short options are used. Reliability and ease of marking and flexibility in terms of tapping into various levels of processing make multiple-choice desirable in listening tests (Elliott & Wilson, 2013); therefore, they were preferred again in the second administration. Furthermore, the item stems were simplified, the tasks did not require any note-taking, and the memory load increased with the difficulty levels of the tasks in the second version. After the second piloting, however, one suggestion from

the TSL instructors regarding memory load was to order the multiple-choice options according to the order of the relevant information in the listening text so that the test takers could follow the options more easily and the memory load imposed upon them would be decreased. This suggestion was valuable since test takers need to carry out many operations at the same time for multiple-choice questions and more emphasis should be put on listening text comprehension and less on memory, reading and other construct-irrelevant variances. This revision, therefore, can be made in the future versions of the test in order to decrease memory load.

Other aspects of context validity in terms of task setting are known criteria, weighting, order of items and time constraints. In both versions of the test, the test takers knew that each correct item was one point and the items in the tasks were all ordered in the order of the information in the listening texts. In the second version, the test writers paid attention to placing adequate space between the parts of the text where the answers of items were located. The spaces increased with the difficulty levels of the tasks due to the density of information and the overall lengths of the texts. However, after the second administration, it was later pointed out by the TSL instructors that the B2 level task had intervals between items longer than estimated (55 seconds). As long intervals can be misleading, one suggestion was that some parts of the text which do not contain any answers or distractors could be removed in the future versions of the test. As discussed above, information about the time constraints were provided to the students in both written and spoken form in the second version. While the test takers were given only 30 seconds to read the questions in the first version, they were given one minute for the A1, A2 and B1 level tasks and three minutes for the B2 level task (due to the length of the items) in the second version. In both versions, the recordings were played only once in this study due to cognitive validity and authenticity concerns. Besides, parameters such as the clarity of speech, redundancy of information, time spaces between the parts that contained answers were controlled to make the speech processable at the designated levels.

Administration Setting

Administration setting as another part of context validity will not be discussed in detail in this study as the tests were administered in actual classroom settings by teachers; however, if this test becomes an institutionalized test, then procedures for test administration will need to be set with the other parties involved in delivering the test.

Task Demands (Linguistic)

Linguistic task demands are related to discourse mode, channel of representation, text length, nature of information, content knowledge and lexical, grammatical and functional resources. During the process of revising the tasks, more attention was given to include a variety of discourse modes in the listening texts. In lower proficiency levels, discourse modes were mostly related with personal environment and topics with immediate relevance; expressive (of individual). In addition to expressive (of individual) discourse mode, exploratory discourse mode can be observed as well in A2 and B1 texts that include dialogues with personal opinions and solutions to problems. In

B2 text in this study, the discourse mode is not about personal matters, but on external issues happening in the outside world, which, in this case, is a festival. Therefore, both informative and exploratory discourse modes are operationalized in this text. These findings indicate that the texts in the test seem to involve a variety of text purposes.

In both versions of the test, only two channels of presentation; i.e. written and spoken, were used in the tasks; the texts were audio recorded and items were printed on the paper. No noise or other distractors were used in the recordings.

Since the listening texts in the first version of the test did not seem satisfactory in terms of text length, a more careful examination was carried out in the second version. The total number of words in the texts and total length of the recordings were calculated in order to see if there is gradation across levels. However, although the listening texts were considered to meet the necessary requirements of the target levels, it was observed after the second administration that the A1 level task had a higher number of words (n=336) than the A2 level task (n=311), which could have created difficulty for the expected level. Since the A1 level task included three short dialogues as opposed to only one in the A2 level task, repetition of some phatic words and phrases for thanking, greeting and taking leave could have increased the word count. Moreover, in terms of text length and delivery speed, the A1 and A2 texts do not seem to be differentiated enough. This can be taken care of either by slowing down the delivery or shortening the text in the A1 level task. Another improvement concerning text length could be made in the B2 level task since it had a considerably higher number of words (n=835) compared to B1 level task (n=389). Although the text was expected to be relatively longer than the others due to its information density, the feedback taken from the TSL instructors after the second administration indicated that some revisions were needed. The situation could be improved by deleting some parts of the text as discussed previously, especially the long intervals between the items.

The nature of information, content knowledge and lexical resources are parameters relating to the vocabulary items used in the tasks. The first version of the tasks included mostly concrete words, which meant revisions were necessary especially for higher-level tasks. For example, the B2 level task included a course announcement, which almost completely consisted of concrete words such as “kurs”, “kayıt”, “program”, “sınıf mevcudu”, “ücret” and so on. In addition, these words could be considered as high frequency words in most academic settings. The C1 level task, on the other hand, included an authentic lecture on information technology; however, similarly, due to its nature of information, mostly concrete words were used in the listening text. Owing to these problems, these texts were not used in the second version and while the new B2 level task was created, special attention was given to add a variety of both concrete and abstract words which are also relatively lower frequency in order to increase the difficulty level of the text. The other texts were also revised to create a better gradation of lexical difficulty across different proficiency levels. During this process, the CEFR Reference Level Descriptors (van Ek and Trim, 1991a, 1991b, 2001) prepared at A1, A2, B1 and B2 levels respectively provided guidance for the test writers. In the second version of the test, at A1 and A2 levels, mostly knowledge of words related to concrete and immediate needs were emphasized. In the corresponding tasks, words related to school, school environment, health, courses, course requirements, etc. such as “doktor” and “yurt” were used in a simple and everyday language.

However, the words "röntgen" or "ilaç" in the A1 level text, as was later realized, caused difficulty for the test takers and considered inappropriate for the target level; therefore, they should be removed from the text in the new version. Apart from these two words, the A1 and A2 level tasks are presumed to comply with the requirements of the proficiency levels according to the CEFR specifications. The texts and items at B1 and B2 level demonstrated a wider range of vocabulary with both concrete and abstract meanings such as "gurur", "emek", "çıkış noktası" and "etkili" as well as some idiomatic phrases and colloquial words. All these indicate that lexical demands of the tasks were improved and made more level-appropriate in the second version.

The tasks were also significantly modified in terms of grammatical difficulty. An analysis of the listening texts and the items in the first version revealed that they did not have the expected levels of difficulty and did not show a satisfactory gradation across levels. Therefore, the sentences in each test task were thoroughly revised according to the requirements of each proficiency level. The Reference Level Descriptors of the CEFR, an analysis of TSL course books and syllabi and opinions of the test writers as native speakers of Turkish helped the revision process. For A1 level only simple sentences, for A2 level mostly simple sentences with the addition of some frequently used coordinate conjunctions, for B1 level a combination of simple and complex sentences with cohesive devices and linkers, and for B2 level mainly relatively longer complex sentences were included in the tasks. However, a further examination of the test tasks after the second administration revealed that the A1 level task included three adverbial clauses, which makes the task unsuitable for the target level and more challenging than the A2 level task and should be altered in the future versions of the test. This finding was also supported with the results from task evaluation questionnaires. Lower-level test takers indicated that the A1 level task (M = 2.34) was more difficult than the A2 level task (M = 2.03) on a four-point scale where "1" means "easy" and "4" means "difficult". The unexpected difficulty level of A1 level task may have resulted from the fact that it may be difficult to detect clauses in Turkish since it is an agglutinating language.

Similar to the lexical and grammatical resources, functional resources required by the test tasks were also examined and underwent changes after the first administration in order to add variation across different proficiency levels. In the second version, a variety of functions were demanded by the tasks such as imparting and seeking factual information, expressing and finding out attitudes, socializing and structuring discourse, and the sub-categories of these functions increased along with the difficulty levels of the tasks.

Task Demands (Interlocutor)

Speech rate is an important consideration in listening tests as it can directly impact on the comprehension level of the test takers. In this study, for a better gradation across tasks, words per minute (wpm) and per second (wps) were calculated in order to see the speech rates after the new texts were created and recorded for the second administration. The calculations showed that there were variations across levels, but the A1 level task seemed to be slightly faster than expected while the B2 level task was slower. The test writers did not consider these two results would have an important effect on the test;

therefore, kept them as they were. However, the statistical analyses carried out after the second administration indicated some problems with the item difficulty levels of the A1 level task. In addition, the task evaluation questionnaires reflected a similar result. The test takers evaluated the A1 level task as being faster than the A2 level task with a mean score of 2.40 in comparison to 2.09. Therefore, some revisions are deemed necessary for the A1 level task. On the other hand, depending on the results of the statistical analysis and task evaluation questionnaires, the speech rate of the B2 level text was not seen as a major problem due to the complex and long messages and an already demanding cognitive and linguistic load in the text. In addition, some pauses between sentences in a long text like this one are thought to be necessary to give it an authentic look and give listeners time to process information; therefore, no significant changes will be made for the B2 level text in terms of speech rate.

Factors related to speakers such as variety of accent, acquaintanceship, number of speakers and gender were also considered to meet the necessary validity requirements. In both versions, no specific accent was used and none of the speakers were known to the test takers. Both male and female speakers were included in the recordings to avoid cultural bias. In the first version, there was more variety in terms of number of speakers such as monologues, dialogues and group discussions whereas in the second version, all the recordings were in the form of dialogues. This might be the only aspect that needs revision related to speakers in order to add variety. Some texts with one or multiple speakers and a lecture from a single speaker, especially at C1 level, may be added to create different contexts and discourse modes in the test.

In conclusion, the examination of the parameters of context validity demonstrates that the second versions of the tasks mostly comply with the required criteria for context validity and provide stronger evidence when compared to their first version. It is also worth noting that some aspects of the A1 level task such as text length, lexical and grammatical resources and speech rate require revisions so as to attain a better gradation across levels. To this end, as discussed earlier, the nature of Turkish as an agglutinating language can require a more meticulous analysis of the language properties of test tasks possibly via using computerized programs or counting syllables per minute instead of words to demonstrate linguistic difficulty. In this way, it will be possible for test writers to create tests in a more standardized way and identify any possible problems related to context validity.

Scoring Validity

The final aspect of the study that needs scrutiny is the scoring validity (reliability) evidence from both administrations of the test. In order to provide reliability evidence, statistical analysis procedures were conducted on the test scores obtained from the test takers in the first and second administrations of the test. Tables 1 and 2 below show the descriptive statistics from the first and second piloting of the test respectively. The most important finding is that the reliability coefficient alpha score is much lower for the overall test in the second administration than it was in the first one. This could have resulted from the much smaller sample size and the lower number of tasks and items in the second administration. However, it should be noted that despite the modifications made on the tasks in terms of cognitive and context validity, the scoring validity of the

test did not improve; on the contrary, it decreased from .954 to .783 and .642.

Table 1. Descriptive statistics of the total test scores in the first administration

N	I_N	Min.	Max.	Mean	SE	SD	Skew.	Kur.	Alpha
55	42	0	39	19.31 45.9%	1.57	11.70	0.11	-1.26	.954

Note. N = Number of test takers, I_N = Total number of items on the test, Skew. = Skewness, Kur. = Kurtosis

Table 2. Descriptive statistics of the test scores in the second administration for the lower-level and higher level test takers

	N	I_N	Min.	Max.	Mean	SE	SD	Skew.	Kur.	Alpha
Lower-level	16	20	6	18	11.44 57.2%	.978	3.91	.40	-1.15	.783
Higher-level	14	22	11	22	17.14 77.9%	.776	2.9	-.604	.487	.642

Note. N = Number of test takers, I_N = Total number of items on the test, Skew. = Skewness, Kur. = Kurtosis

In addition to the overall reliability of the test, more detailed reliability evidence comes from individual item analyses. The item mean (IF), item discrimination (ID, Corrected-Item Total Correlation-CITC) and alpha if item deleted (AIID) values need to be considered to demonstrate the extent to which the items function well in the test. The results indicated that a few items were problematic. However, these statistics were ignored because the tasks went through extensive revision for the second administration. In the second administration, however, there were a higher number of items with problematic results. Certain items had a negative effect on the overall reliability of the test. The item statistics for the higher-level test takers showed that most of the items had very low discrimination values (CITC) and had a lowering effect on the alpha coefficient. The reasons for this were investigated through the analysis of test takers' responses for further improvement of the test; however, detailed analyses cannot be reported here due to space limitations. Readers are referred to Tozlu (2017) for the details. Nevertheless, our contention is that an important reason behind this could again be the very small sample size of the higher-level test takers and their homogeneity. Therefore, conducting the analyses for the results obtained from a greater number of test takers can yield different results.

Along with reliability values, we can also look at the mean scores of the tasks in order to see if the tasks are ordered in accordance with their target proficiency levels. Table 3 below demonstrates that in the first piloting the tasks failed to comply with the required difficulty levels of the target tasks.

Table 3. Mean scores for the tasks in the first administration

Task	Mean scores	Mean scores out of 100
A1 level task	2.96 /6	49.33
A2 level task	3.31 /6	55.1
B1 level task	5.96 /10	59.6
B2 level task	3.33 /10	33.3
C1 level task	3.75 /10	37.5

Table 4 shows that in the second piloting, the A1 level task still seems to be problematic as it has a much higher difficulty level than expected for the lower-level test takers. The reason for this situation can be linked to the cognitive and contextual requirements of the A1 level task, which also need improvements pertaining to context validity parameters such as text length, lexical and grammatical resources and speech rate as discussed above. On the other hand, the difficulty of the tasks for the higher level test takers conform to our expectations (see Table 4).

Table 4. Mean scores for the tasks in the second administration for the lower-level and higher level test takers

	Task	Mean scores	Mean scores out of 100
Lower-level	A1 level task	3.13/6	52.16
	A2 level task	5.25/8	65.62
	B1 level task	2.88/6	48
Higher-Level	A2 level task	7.50/8	93.75
	B1 level task	4.71/6	78.5
	B2 level task	4.93/8	61.62

Besides, when all the individual items across tasks were ordered in terms of their item facility values, most of the items were aligned with the corresponding tasks conforming to the expected difficulty of the tasks; e.g. most B1 items having lower item facility than B2 items. Therefore, despite some problems detected for the A1 level task as a result of the examination of context validity, we can argue that the changes made in the first versions of the tasks related to cognitive and context validity requirements, especially linguistic task demands, had a positive effect on the ordering of the items in terms of difficulty and produced more satisfactory results.

Conclusion

In this study, we aimed to show the steps taken to ensure cognitive, contextual and scoring validity of a TSL Academic Listening Test and the impact of the modifications made on the first version of the test in accordance with the requirements of Weir's (2005) socio-cognitive framework for test validation. To this end, we showed the

aspects of the tests which were modified for certain purposes and the effects these changes had on the second administration of the test. After the theoretical and statistical analyses of the second version of the test tasks, it was observed that in terms of cognitive validity the test tasks succeeded in conforming to the requirements of the target CEFR levels and showing a better alignment with Field's (2013) and Weir's (1993) frameworks for listening. The contextual features of the tasks were also considered to be mainly satisfactory in terms of context validity parameters. On that note, one suggestion regarding establishing context validity would be to develop more materials such as TSL course materials, textbooks, tools or software programs based on current views of language learning and language use in order to help test writers set a more standardized way of controlling language features of tests for TSL learners. With the help of such standardized tools, validation studies of TSL tests will be more systematic.

As for scoring validity (reliability), the evidence gathered from the test results in the second administration indicated that the reliability of the test tasks and the whole test decreased when compared to the first administration although there were improvements regarding difficulty levels of the tasks. This leads to an important conclusion in this study; that is, with high reliability a test may still have misaligned, poorly calibrated items that may not reflect the theoretical cline of subskills. On the other hand, a good transference of designated cognitive skills and contextual demands from theory to test tasks does not guarantee item quality. Items should then be checked again for technical problems and the assumptions of statistical analysis such as sample size, heterogeneity and normal distribution should be maintained. In this study, one of the limitations was the sample size in the second administration of the test. The relatively worse reliability scores could have resulted from the small number of participants. In addition, the homogeneity of the language levels within the two groups of participants in the second piloting could have had a negative impact on the reliability coefficients. Besides, it is advisable to use more advanced statistical techniques such as Item Response Theory to be able to do sample-independent analysis. The test tasks can also be administered to a higher number of test takers with more heterogeneous backgrounds to gain more conclusive results for reliability.

References

ALTE, (2011). Manual for language test development and examining. Downloaded from https://www.alte.org/resources/Documents/ManualLanguageTest-Alte2011_EN.pdf

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

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

Borsboom, D., Mellenbergh, G. J., & Heerden, J. V. (2004). The concept of validity. *Psychological Review, 111*(4), 1061-1071.

Council of Europe, (2001). *Common European framework of reference for languages: Learning, teaching and assessment.* Cambridge: Cambridge University Press.

EALTA guidelines for good practice in language testing and assessment. Downloaded from http://www.ealta.eu.org/documents/archive/guidelines/English.pdf

Elliott, M., & Wilson, J. (2013). Context validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening*, Studies in language testing, 35 (pp.152-241). Cambridge: Cambridge University Press.

Field, J. (2013). Cognitive Validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening*, Studies in language testing, 35 (pp.77-151). Cambridge: Cambridge University Press.

Geranpayeh, A. (2013). Scoring validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening*. Studies in language testing, 35 (pp.242-272). Cambridge: Cambridge University Press.

Kane, M. T. (2013) Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement 50*(1), 1–73.

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.

Knoch, U., & Elder, C. (2013). A framework for validating post-entry language assessments (PELAs). *Papers in Language Testing and Assessment, 2*(2), 48-66.

McNamara, T. (2000). *Language testing.* Oxford: Oxford University Press.

Messick, S. A. (1989). Validity. In R. L. Linn (Ed) *Educational measurement* 13-103. New York. American Council on Education. Mac Millian Publishing Company.

Richards, J. C. (1983). Listening comprehension: Approach, design, procedure. *TESOL quarterly*, *17*(2), 219-240.

Tozlu, E. (2017). *The development of a listening test for learners of Turkish as a foreign language* (Unpublished master thesis). Boğaziçi University, İstanbul, Turkey.

Trim, J. L. M. (2009). *Breakthrough*. Retrieved from https://www.coe.int/t/dg4/linguistic/Source/FinalBreakthrough%20specification_6Nov01.rtf

Van Ek, J., & Trim, J. L. M. (1991a). *Threshold 1990*. Cambridge: Cambridge University Press.

Van Ek, J., & Trim, J. L. M. (1991b). *Waystage 1990*. Cambridge: Cambridge University Press.

Van Ek, J., & Trim, J. L. M. (2001). *Vantage*. Cambridge: Cambridge University Press.

Walt, J. L., & Steyn, F. (2008). The validation of language tests. *Stellenbosch Papers in Linguistics, 38*, 191-204.

Weir, C. J. (1993). *Understanding and developing language tests*. New York & Toronto: Prentice-Hall.

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

Young, J.W., So, Y., & Ockey, G.J. (2013). *Guidelines for best test development practices to ensure validity and fairness for international English language*

proficiency assessments. Educational Testing Service. https://www.ets.org/s/about/pdf/best_practices_ensure_validity_fairness_english_language_assessments.pdf

İkinci Dil Olarak Türkçede Dinleme Becerisinin Ölçülmesinde Güvenirlik ve Geçerlik Dengesinin Sağlanması

Özet

Bir değerlendirme aracından elde edilen sınav sonuçlarına dayanarak aldığımız kararları haklı göstermek için, o değerlendirme aracının etkinliliğine dair kanıtlara sahip olmamız gereklidir. Geçerlik kanıtı, hem sınavın uygulanmasından önce hem de sonra pek çok kaynaktan toplanabilir. Bu çalışmada, İkinci Dil olarak Türkçe Akademik Dinleme Testi'nin etkililiğini göstermek amacıyla toplanan pek çok türde geçerlilik kanıtı sunulmuştur. Çalışma, bilişsel, bağlamsal ve puanlama geçerliği (güvenilirliği) türlerinde, sınavın birinci ve ikinci versiyonlarından toplanan kanıtları sunmakta ve sınavın birinci uygulamasından sonra yapılan değişikliklerin, sınavın bilişsel ve bağlamsal geçerlilik ve güvenilirlik iddialarına olumlu bir etkisi olup olmadığını incelemektedir. Araştırma sonuçları, sınavın birinci versiyonunda yapılan değişikliklerin bilişsel ve bağlamsal gereksinimler açısından geçerlilik iddialarını güçlendirmiş olduğunu ancak sınavın güvenilirlik puanları ikinci versiyonda kötüleştirdiğini göstermiştir. Bu da bize bir sınavın, gerekli bağlamsal ve bilişsel özellikleri tümüyle yansıtarak, sağlam temeller üstüne kuruması gerektiğini ama bunun teknik kaliteyi tamamıyla garanti etmesinin mümkün olmadığını göstermiştir. Sınavın puanlama geçerliliğinin (güvenirlik) de titizlikle sağlanması gereklidir.

Anahtar Kelimeler: Bilişsel geçerlilik, bağlamsal geçerlilik, puanlama geçerliliği, güvenilirlik, dinleme becerisinin değerlendirilmesi, Türkçenin ikinci bir dil olarak değerlendirilmesi

Appendix

Table 1. Lower-level Test Takers' Perceptions of Cognitive Processes in A1 Level Task (n=16)

In order to answer this question correctly I had to.....	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1. understand specific bits of information in the dialogue	11*	13*	13*	13*	9*	9*
2. understand just the main idea(s)	6*	3	3	3	7*	6*
3. understand the details used to explain the main idea(s)	4	3	3	3	5	4
4. differentiate between important and less important information	11*	10*	6*	6*	7*	6*
5. understand what the dialogue is about briefly	9*	6*	6*	5	8*	6*
6. understand how information in the whole dialogue fits together	3	4	3	3	7*	4
7. pay attention to the speakers' attitude and tone	2	2	3	2	4	3
8. understand what the speaker's intention is when using a certain sentence	3	3	1	1	2	3
9. rely on my general world knowledge	3	1	0	0	4	6*

Note: Asterisks indicate the reading operations taken as prominent in the execution of the tasks.

Table 2. Lower-level (n=16) and Higher-level (n=14) Test Takers' Perceptions of Cognitive Processes in A2 Level Task

In order to answer this question correctly I had to.....		Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8
1. understand specific bits of information in the dialogue	Lower-level	11*	15*	16*	16*	15*	15*	14*	14*
	Higher-level	7*	8*	12*	12*	11*	11*	10*	14*
2. understand just the main idea(s)	Lower-level	7*	2	2	1	1	2	2	1
	Higher-level	2	2	0	1	0	0	1	2
3. understand the details used to explain the main idea(s)	Lower-level	4	4	4	4	3	5	7*	7*
	Higher-level	5*	5*	6*	5*	6*	6*	6*	6*
4. differentiate between important and less important information	Lower-level	9*	7*	9*	8*	7*	6*	7*	6*
	Higher-level	4	3	5*	5*	3	4	3	3
5. understand what the dialogue is about briefly	Lower-level	10*	5	6*	5	7*	4	4	3
	Higher-level	7*	1	2	2	2	3	2	2
6. understand how information in the whole dialogue fits together	Lower-level	4	0	0	2	1	1	3	1
	Higher-level	3	2	4	2	4	1	4	1
7. pay attention to the speakers' attitude and tone	Lower-level	6*	2	2	2	1	1	2	1
	Higher-level	0	0	0	0	1	1	0	0
8. understand what the speaker's intention is when using a certain sentence	Lower-level	3	2	2	2	1	1	2	1
	Higher-level	1	0	0	1	1	0	0	0
9. rely on my general world knowledge	Lower-level	5	1	2	2	3	1	1	1
	Higher-level	1	1	2	1	1	1	1	1

Note: Asterisks indicate the reading operations taken as prominent in the execution of the tasks.

Table 3. Lower-level (n=16) and Higher-level (n=14) Test Takers' Perceptions of Cognitive Processes in B1 Level Task

In order to answer this question correctly I had to...		Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1. understand specific bits of information in the dialogue	Lower-level	14*	13*	12*	11*	10*	13*
	Higher-level	7*	11*	8*	9*	6*	6*
2. understand just the main idea(s)	Lower-level	4	3	0	1	4	4
	Higher-level	3	1	3	1	2	3
3. understand the details used to explain the main idea(s)	Lower-level	6*	8*	7*	8*	7*	6*
	Higher-level	5*	4*	4*	3	4*	2
4. differentiate between important and less important information	Lower-level	10*	8*	8*	8*	8*	6*
	Higher-level	3	7*	4*	4*	4*	3
5. understand what the dialogue is about briefly	Lower-level	5	3	2	4	3	1
	Higher-level	2	3	1	3	2	3
6. understand how information in the whole dialogue fits together	Lower-level	5*	5*	6*	6*	6*	6*
	Higher-level	3	4	3	3	3	2
7. pay attention to the speakers' attitude and tone.	Lower-level	5	0	1	2	3	4
	Higher-level	2	2	2	1	2	3
8. make an inference based on the information in the text	Lower-level	6*	2	3	3	5*	3
	Higher-level	2	1	1	2	4*	3
9. understand relations between the speakers and the situation they are in	Lower-level	1	2	3	2	3	2
	Higher-level	1	1	1	2	2	1
10. understand what the speaker's intention is when using a certain sentence	Lower-level	2	0	1	2	1	4
	Higher-level	2	0	0	1	1	1
11. understand what an unknown word/phrase means based on the information in the text	Lower-level	0	0	0	1	2	1
	Higher-level	1	1	1	2	1	2
12. rely on my general world knowledge.	Lower-level	3	4	3	3	4	5
	Higher-level	1	1	1	0	0	1

Note: Asterisks indicate the reading operations taken as prominent in the execution of the tasks.

Table 4. Higher-level Test Takers' Perceptions of Cognitive Processes in B2 Level Task (n=14)

In order to answer this question correctly I had to...	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8
1. understand specific bits of information in the dialogue	8*	7*	8*	7*	5*	5*	5*	3
2. understand just the main idea(s)	3	3	2	2	3	2	2	2
3. understand the details used to explain the main idea(s)	5*	5*	5*	5*	4	4	6*	7*
4. differentiate between important and less important information	6*	5*	8*	7*	7*	6*	6*	5*
5. understand what the dialogue is about briefly	5*	3	4	4	2	2	8*	5*
6. pay attention to the speakers' attitude and tone	1	1	1	1	1	1	3	1
7. understand how information in the whole dialogue fits together	1	1	1	2	2	3	2	7
8. understand how certain parts are linked to others in the dialogue	5*	4	5*	5*	6*	6*	6*	8*
9. make an inference based on the information in the text	2	2	3	3	5*	6*	4	7*
10. understand what the speaker's intention is when using a certain sentence	0	0	0	0	1	1	1	0
11. understand what an unknown word/phrase means based on the information in the text	2	1	2	1	1	2	2	1
12. rely on my general world knowledge.	3	3	2	2	2	2	2	2

Note: Asterisks indicate the reading operations taken as prominent in the execution of the tasks.