Score Dependability of a Speaking Test of Turkish as a Second Language: A Generalizability Study

Talip Gülle and Gülcan Erçetin

Abstract

Test developers have to attend to all aspects of validity throughout test development and implementation. As one of the major aspects, scoring validity has to be established for the dependability of scores assigned to a test performance. This study is an investigation into the scoring validity of a speaking test of Turkish as a second language (TSL). For this purpose, in this study, six tasks and a rating scale were developed and administered to twenty-four L2 learners of Turkish whose performance was evaluated by four raters. The score dependability was investigated through Generalizability (G) and Decision (D) analyses. The results indicated that most of the score variation could be attributed to test takers, and not to error variance, i.e. raters and tasks.

Key words: Assessing speaking in Turkish as a second language, scoring validity, generalizability analysis

Introduction

This study presents an investigation into the scoring validity of a speaking test designed as part of an academic language test consisting of four skills, which aimed to assess Turkish as a second language (TSL). The investigation into reading, listening and writing skills are presented in the other chapters in this volume. The development and validation-related work of the speaking component of the test is presented in detail in Gülle (2015), and, in this article the focus will be specifically on one component of the validation process, namely scoring validity.

Test development and validation is a process that is complex, continuous and challenging and should be carried out in a systematic way (Hasselgreen, 2004). The concept of validity relates to score interpretations and use (Messick, 1989) that need to be supported by theoretical and empirical evidence collected in a wide range from the initial stages of test design to the implementation of a test as well as to the consequences that are produced by the test whether for the individual test-taker or the society at large. It is essential, then, that tests go through a rigorous validation process.

The current study is informed mainly by Weir's (2005a) socio-cognitive framework for test development and validation in relation to the assessment of L2 speaking ability (Khalifa & Weir, 2009; Shaw & Weir, 2007; Taylor, 2011). In Weir's socio-cognitive framework, validity is a construct with multiple components, i.e. context validity, test-taker characteristics, theory-based validity, scoring validity, consequential validity and criterion-related validity, each of which is a sine qua non for establishing that the test serves the purpose it is intended for. Thus, Weir's (2005a) framework addresses a number of questions in relation to the physical/physiological, psychological and experiential characteristics of *test-takers* who are targeted by the test; the

Talip Gülle, Boğazici University, Department of Foreign Language Education, talipgulle@gmail.com
Gülcan Erçetin, Boğazici University, Department of Foreign Language Education, gulcan.ercetin@gmail.com

appropriateness of the test tasks in terms of *cognitive processes* required; *task features*, *task appropriacy*, and *fairness* in tasks; *score dependability*; *consequential impacts* of the test; and lastly, *external evidence* for construct validity (Taylor, 2011). As a thorough discussion of each aspect of validity is beyond the scope of the present paper, and consistency and dependability of scoring in performance-based assessment is a core issue, only scoring validity will be discussed and examined here in relation to the assessment of TSL.

Scoring Validity

The phrase *scoring validity* was first adopted by Weir (2005a) to cover all aspects of the assessment process which may influence the consistency and dependability of scores (Taylor & Galaczi, 2011). In other words, it was coined as an umbrella term for all aspects of reliability. As Shaw and Weir (2007) suggest, scoring validity indicates whether test scores are based upon appropriate criteria, exhibit agreement in marking, are free from measurement error, are consistent over time and in terms of content sampling and allow reliable decision-making (p. 143).

A number of factors may threaten the scoring validity of an assessment, including imprecise scaling descriptors, inconsistent raters or unsuitable rating procedures, inadequate training or unsystematic grading (Shaw & Weir, 2007). Additionally, a limited range of tasks and interaction types, allocation of different amounts of time and weight to different components of the test, variation in terms of channels of communication, variation that stem from the interlocutor, imprecise or ambiguous rubrics or inconsistency across different administrations of the test can threaten the scoring validity of a speaking test (Taylor & Galaczi, 2011). As such, score dependability, and hence validity should be ensured not only in terms of rater variables which are typically expressed in the intra- and inter-rater correlation coefficients but also in relation to the following parameters: assessment criteria/ rating scale, rating process, rating conditions, rater characteristics, and rater training.

Assessment Criteria and Rating Scales

Rating scales are an essential part of the assessment of productive skills, that is, writing and speaking (Alderson, Clapham & Wall, 1995; Bachman & Palmer, 1996; McNamara, 1996) because they delimit the inferences to be made based on test scores and meanings attributed to test scores. Since rating scales are based on construct definitions (Fulcher, 2003), their development requires describing complex phenomena "in a small number of words on the basis of incomplete theory" (North, 2000 as cited in Luoma, 2004, p. 13). In other words, the inadequacy of solid evidence about language learning poses challenges in summarizing descriptors into brief statements (Luoma, 2004).

Rating scales have been divided into two main types as behavior-oriented and construct-oriented. The former adopts a "real-world" approach (Bachman, 1990, p. 344) and defines language ability in relation to real life performance, focusing on language use in specific contexts, whereas the latter defines language ability independently of the context of language use and is based on a theoretical model of the construct that is being assessed. Behavior-oriented scales have been subject to criticism on the basis that it is not possible to generalize from test scores to speaking situations that are not captured in test tasks, and that they are inadequate in providing information about specific components of language ability (Bachman & Savignon, 1986; Bahman & Palmer, 1996). Construct oriented scales, on the other hand, allow for generalization beyond the tasks due to their focus on constructs rather than test tasks (Bachman, 1990; Bachman & Palmer, 1996).

Another distinction between scales relates to the scoring approach taken up by a rating scale: holistic vs. analytic. Holistic scoring involves impressionistic assessment of a test taker's ability using a single rating scale (Davies et al., 1999), which yields practical and relatively quick assessement (Luoma, 2004). Additionally, holistic descriptors can provide a summary of skill levels that can be intuitively accessed and used without requiring any specialist knowledge of linguistics or language assessment (Galaczi & ffrench, 2011). However, these scales are insufficient in identifying the strengths and weaknesses of test taker performances (Luoma, 2004) and, due to their intuitive nature, they lack empirical foundation (Weir, 1993). As such, raters may focus on different aspects of test taker performance in arriving at judgments using self-generated criteria (Weir, 1993; Barkaoui, 2007).

Analytical scales, on the other hand, require raters to attend to a number of criteria such as fluency and coherence, lexical resource, grammatical range and accuracy, pronunciation (International English Language Testing System, 2009); grammatical resource, lexical resource, interactive communication, discourse management, pronunciation (University of Cambridge ESOL Examinations, 2015); delivery, language use, topic development (Educational Testing Service, 2004); range, accuracy, fluency, interaction, coherence (Council of Europe, 2001). Raters provide a score for each category. Analytic scales have advantages such as providing detailed guidance to the raters (Luoma, 2004), which can contribute to rater agreement and rater reliability unless the criteria are inexplicit or vague (Weir, 1993). Additionally, they draw the raters' attention to aspects of performance that may be overlooked otherwise and provide information on specific strengths and weaknesses of test takers by taking account of different levels of ability in subskills (Hughes, 2003).

A number of disadvantages of analytic scales are also identified (Galaczi & ffrench, 2011). In the first place, the rating criteria may not be functioning independently, or raters may not distinguish between them. In addition, it is time consuming and may not be practical in certain assessment contexts, such as when the rater is at the same time the interlocutor in a test. Luoma (2004) especially warns that making multiple judgments may increase cognitive load on raters. In the Common European Framework of Reference (CEFR) guidelines, the psychological upper limit for raters is suggested to be seven categories (Council of Europe, 2001).

Apart from the decision as to what type of a scale is to be employed, rating scale development also requires decisions regarding the number of levels and the number of criteria to be used in the scale and descriptions of score meanings (Luoma, 2004). Luoma (2004) points out that although more levels can provide more detailed information, care needs to be taken to ensure that the raters are able to distinguish between the levels consistently. Moreover, the criteria should be conceptually independent. She suggests that five to six criteria may be close to the maximum. With regards to the number of levels, Isaacs and Thomson (2013) noted that while 5-point scales were too constraining for some raters, 9-point scales were difficult for them as they would be unable to differentiate between the levels. In addition to having a justifiable number of levels and criteria, the scale descriptors should be concrete, practical, and easy for raters to memorize (Luoma, 2004). Fulcher (1996a) criticizes intuitive construction of scales, and instead favors scale development based on empirical performance data.

Rating Process and Rating Conditions

Producing an appropriate rating scale does not ensure score reliability since the conditions under which raters perform may also influence score variability. The rating process and rating condition parameters relate to the decision-making processes that raters undergo as well as to the conditions under which raters operate while scoring performances (Taylor & Galaczi, 2011). Taylor and Galaczi (2011) argue that these parameters naturally differ and affect the rating process when the raters have to make real-time judgments as compared to rating video-taped performances, in which case they would have the opportunity to replay the video if they needed to. Similarly, they point out that a rater who participates as an interlocutor in the testing process would have more cognitive load in comparison to one who does not. Along with the efforts that are made so that conditions are optimized for raters in practical testing situations, rater training is also indispensable in the standardization of the testing process.

Rater Characteristics and Rater Training

Research has shown extensive variability in test scores that are attributable to rater effects. Significant effects of bias towards specific test takers (Kondo-Brown, 2002; Lynch & McNamara, 1998), specific test tasks (Lynch & McNamara, 1998; Wigglesworth, 1993) and specific rating criteria (Eckes, 2009; Wigglesworth, 1993) have been observed. Rater effects such as severity or leniency, halo, central tendency, randomness, and bias are viewed as sources of systematic error variance (Eckes, 2005; Hoyt, 2000; Myford & Wolfe, 2003, 2004). Substantial variation in rater severity has been noted in various studies (Chalhoub-Daville & Wigglesworth, 2005; McNamara & Adams, 1991; Weigle, 1998) despite extensive rater training (Eckes, 2009; Lumley & McNamara, 1995; Weigle, 1998).

Specifically, previous research has shown that non-native speakers are more severe in their assessment of writing and speaking performances than native speakers (e.g., Shi, 2001). Similarly, raters' familiarity with the test takers' first language may also have an impact on their evaluation of learners' oral performances (Carey, Mannell, & Dunn, 2011; Chalhoub-Deville & Wigglesworth, 2005). On the other hand, these observations are not consistent. For instance, Kim (2009) and Zhang and Elder (2011) observed only marginal differences in the severity of native and non-native raters although both noted that native speaker raters were more detailed in their comments. Brown (1995) found little evidence that native speakers would be more suitable than nonnative speakers, suggesting that raters from different languages can be trained to be as effective as native speaker raters, which was also corroborated by later research

(Johnson and Lim, 2009; Xi & Mollaun, 2009, 2011). Winke, Gass and Myford (2013) found that L1 familiarity contributed to rater leniency, but the effect sizes were small.

McNamara (1996) suggests that rater variability is a “fact of life”, adding that it may be unachievable to eradicate differences between raters (p. 127). Score variation due to raters can be reduced through rater training (Elder, Knoch, Barkhuizen & von Randow, 2005; Johnson & Lim, 2009; Weigle, 1994; Xi & Mollaun, 2009, 2011). Statistically, tools such as Many-Facet Rasch Measurement can adjust score variance produced by assessment conditions (see Eckes, 2011 for an introduction to Many-Facet Rasch Measurement). In addition, Generalizability and Decision analyses can be used to improve the assessment design by investigating the ideal number of raters for score reliability. The following section will provide a brief introduction to Generalizability and Decision analyses since these analyses were utilized to investigate score dependability in the current study.

Score Dependability and Generalizability Theory

In measuring proficiency in an L2 and linking observations of test performances to interpretations about a test taker’s ability to use a language in a particular context or for a particular purpose, language testers aim to be able to generalize beyond the test. These inferences and generalizations are closely connected to issues of validity and dependability, i.e. reliability[2] of test scores (Deville & Chalhoub-Deville, 2006). Chalhoub-Deville (2006) writes, “dependability or reliability, in a broad sense, refers to the consistencies of data, scores, or observations obtained using elicitation instruments” (p. 2) and lists a number of sources of error that can limit the degree of reliability, such as the elicitation method, the number of elicitations, and the influence of the interlocutor or observer involved in the elicitation.

Psychological and educational research has examined the issue of reliability predominantly through classical test theory (CTT), which partitions observed-score variance into systematic variance (also called true score variance) and random variance (also called error variance) (Feldt & Brennan, 1989, cited in Brennan, 2000). In other words, a classical reliability coefficient generally implies a single undifferentiated source of measurement error. Therefore, when the assessment involves a single measurement facet, CTT provides sufficient information regarding the generalizability of test scores (Lee, 2006). However, the usefulness of CTT “depends on the researcher’s ability to estimate true score and error variances from data. With practical application of CT[3], we find that error variance is not a monolithic construct; error arises from multiple sources” (Shavelson, Webb, & Rowley, 1989, p. 922). Although the reliability coefficient is the most widely used indicator of reliability, it may not always provide the most appropriate estimate, since it may under- or over-estimate reliability depending on how an assessment system is constructed (Deville & Chalhoub-Deville, 2006).

Since assessment of productive skills such as speaking involve more than one one random facet (Lee, 2006), the assessment context introduces a range of factors that influence the performance of a test taker (Lynch & McNamara, 1998) other than the object of measurement (i.e., test takers). Of these factors, variability in assessment tasks

[2] The terms *reliability* and *dependability* will be used synonymously in the following sections.
[3] Shavelson, Webb and Rowley (1989) uses CT as the abbreviation of Classical Test Theory.

and rater judgements as random facets have been widely researched (e.g., Chalhoub-Daville & Wigglesworth, 2005; Lynch & McNamara, 1998; Wigglesworth, 1993) through generalizability theory (G theory), especially in relation to speaking (Bachman, Lynch & Mason, 1995; Brown & Ahn, 2011; Lee, 2005, 2006; Lee & Kantor, 2005, Lynch & McNamara, 1998; Sawaki, 2007). Brennan (2000) notes "generalizability (G) theory liberalizes CTT by providing models and methods that allow an investigator to disentangle multiple sources of error that contribute to E[4]" (p. 339). Shavelson, Webb and Rowley (1989) list a number of ways in which G-theory extends the framework of CTT. These include (a) attending to more than one source of measurement error, analyzing each source on its own and offering ways to maximize reliability; (b) estimating the magnitude of each source of error variance; and (c) allowing test producers to design a test where error variance due to particular sources can be minimized through Decision (D) studies.

Briefly, G-theory employs the analysis of variance (ANOVA) to partition the variation in scores into different sources and the interactions between them (Huang, 2012). These sources of variance, also called variance components (VCs), are then used to estimate the impact of various changes in assessment design (generally different numbers of raters and tasks in the case of speaking assessment) on the generalizability or dependability (which is analogous to reliability in CTT) of the scores (Brown, 1999; Brown & Kondo-Brown, 2012) through a decision study (D-study). So, as Huang (2012) notes, a D-study uses the same data as the G-study and provides estimations about the relative effects of specified numbers of conditions for each VC (For example, for using one task and two raters, two tasks and one rater, two tasks and two raters, etc.).

Given the lack of research on the assessment of speaking ability in TSL, the current study aims to analyze the factors that may have an impact upon scoring validity with a view to substantiating the validation arguments for the dependability of scores. The following research questions are addressed in relation to the speaking test developed to assess speaking in TSL:

1. How reliable is the rating process with the newly developed Rating Scale (see appendix)?
2. How dependable are the scores assigned to the test takers?
   2.1. To what extent are the analytic scores dependable?
   2.2. How many tasks and raters would produce relatively more dependable scores?

Methodology

Six tasks were designed based on CEFR 'can do' statements (for a full discussion of task development process, see Gülle, 2015). A rating scale was also developed to assess the performance of the test takers on these tasks (see the appendix for Turkish versions of the scales and Gülle, 2015 for the English). With respect to the reliability of the rating *process*, the aspects of the test that relate to the raters and the rating process were analyzed. The raters, rating procedures and rating conditions are described for the standardization of the rating process. A Generalizability analysis was employed to

[4] E is the undifferentiated random error term.

examine the dependability of the scores assigned by the raters, and Decision studies were used to determine the number of tasks and raters that would produce more dependable scores.

Participants

The participants were 22 learners of Turkish (16 female, 6 male) enrolled in Turkish for Foreigners classes at a university in Turkey. Their age ranged from 20 to 37 (*M* = 21). They were at that university for either one or two semesters through international student exchange programs. Most of them had started learning Turkish for over two years before the data collection, and only five of them had been learning Turkish for less than two years. The participants' average length of residence in Turkey was relatively short (15 months), with most of them living in Turkey for about three months at the time of data collection.

Speaking Tasks

The six speaking tasks used in the present study were developed by one of the researchers with guidance from the course instructors and experts. Since the test was intended to be used as a general proficiency test, test tasks were informed by the CEFR 'can do' statements from A2 to C2 levels. The test development process included various stages such as selection of task types, writing of task items, consulting with experts and a pilot test. The data reported for the present study is based, not on the initial pilot test, but on the second testing, which included the revised forms of the initial tasks (with the addition of one new task) following student feedback, expert feedback and analysis of the first pilot test results. Four of the six tasks were individual tasks where the examiner gave the instructions and asked the questions, and there were two other tasks which required the interaction of two participants with each other. Table 1 provides the CEFR levels aimed by and the time allotted to each task. The amount of time to be allotted to the tasks was decided based on the average amount of time that the tasks took to complete in the first pilot examination and test taker feedback.

Table 1. Duration of the speaking tasks

Task	Intended level	Time given for the task
1	A2 to C2	7 min.
2	B1	4 min.
3	B2 and C1	5 min.
4	A2 and B1	4 min.
5	B2 to C2	4 min (1 min. for preparation; 3 min. for speaking)
6	A2 and B1	5 min.

Data Collection Procedures

Since the student participants had a tight schedule and participation was voluntary, the six tasks were divided into two sets, Set A (Task 1, 2, 3) and Set B (Task 4, 5, 6). Both sets included two individual tasks and one paired task. Twelve of the participants took one set of the test, and 12 others took the other set. The participants were informed about the test and a short description of each task was sent to the participants via e-mail before the testing session. Three examiners were trained about how to implement each task in order to ensure standardization. All of the participants were audio recorded to be rated afterwards. The tasks were administered over a period of three weeks since the administrations had to be scheduled in the time slots that suited the programs of test takers.

Rating Scale

The rating scale used in the current study was based on level descriptors selected from the CEFR, as the task development also was informed by the CEFR (see appendix) The rating scale included Fluency (F), Grammatical Range and Accuracy (A), Lexical Resource (L), Coherence (C), and Interaction (I) dimensions. It should be noted that the test takers were scored on ‘Interaction’ in Task 3 and Task 6, which required interaction among the test takers, and they were scored on ‘Coherence’ in the other tasks. Thus, in both Set A and Set B, in the third tasks which involved interaction, Coherence was replaced with Interaction; while the other three dimensions (i.e., fluency, grammatical range and accuracy, and lexical resource) were used in all tasks. The use of Interaction rubrics in tasks that require test taker-test taker interaction, and Coherence rubrics in individual test-taker performances is also discussed by Luoma (2004). In her discussion of the adaption of the CEFR speaking scales for specific speaking assessment contexts, Luoma (2004) points out that “The interaction scale provides some concrete suggestions for wordings when rating interactive skills, while for tasks that require long turns by a single speaker the Coherence scale may provide some useful concepts” (p. 71). The Coherence scale in the current study relates to the logical organization, development and connection of ideas within the discourse produced by the individual test taker; the Interaction scale, on the other hand, is about a process of joint construction, in which the test taker has to relate his/her utterances to a shared discourse framework. The organization, development and connection of ideas in the Interaction scale are assessed in relation to the co-constructed discourse; and in this respect the Interaction rubrics incorporate aspects of the Coherence rubrics.

Rating Procedure

For the generalizability and decision analyses, four raters scored the performance of each candidate on each of the four dimensions included in the rating scale on all three tasks (12 test takers took the three tasks in Set A and 12 others took the other three tasks in Set B).

The raters were native speakers of Turkish. An intensive rating training session on the rating scale and scoring procedures took place. Student performances from the first and second pilot tests were used to exemplify performances at different levels for each dimension in the rating scale (i.e., fluency, grammatical range and accuracy, lexical resource, coherence/interaction). After the ratings, the scores were compared by the researcher and when discrepancies were observed, the raters were asked to rate for a second time these particular performances on which there were disagreements. The raters were free to change the scores they assigned in the first rating or give the same score again. In few cases, the raters decided to keep the initial scores, while in most others they assigned a different score. The goal in the rater training and monitoring was to minimize the differences in the ratings.

Data Analysis

With regard to the first research question, the following parameters related to scoring validity were analyzed so as to examine the reliability of the rating process: assessment criteria/rating scale, rating process, rating conditions, rater characteristics, and rater training (Taylor & Galaczi, 2011). The rating scale and assessment criteria are presented in Appendix.

The second research question was addressed through Generalizability (G study) and Decision (D study) studies. These analyses were carried out on the EduG-6e software program. The aim of a G study is to partition the variation observed in test scores to reveal the sources of variation (e.g., persons, raters, tasks) and their interactions through the analysis of variance (ANOVA), which calculates the contribution of each source of variance to the overall score variance. The aim of a D study, on the other hand, is to predict the relative effects of specified numbers of conditions (for example, using one rater and three tasks or two raters and two tasks, etc.) on score dependability by using the data obtained from the G-study. Note that since the Coherence dimension was replaced with the Interaction dimension in tasks that required test taker-test taker interaction, for the G and D studies, Coherence and Interaction scores were pooled into the coherence/interaction (C/I) scale.

These analyses consisted of two phases – the first phase targeted overall scores of the three tasks in each set and the second phase targeted each task separately. In each phase both G and D studies were conducted. The first phase featured a fully crossed p x t x r design (persons crossed with tasks crossed with raters) for the scoring dimensions for each set, Set A and Set B, separately. A p x t x r design means that *each* test taker took *each* task and was rated by *each* rater on *each* task. In this phase, the analyses were based on averaged variance components of all three tasks. The second phase featured a fully crossed p x r design (persons crossed with raters) for each scoring dimension on each task. In other words, the relative contributions of persons and raters to overall score variance was investigated, and dependability coefficients for analytic scores at individual task level was obtained so as to determine which tasks were likely to introduce more error variance in rating a particular scoring dimension.

In the D studies in each phase, the numbers of the facets (that is, tasks and raters) were varied to examine their impact on the phi coefficients (dependability) of the analytic scores. In other words, D studies showed how score dependability would change when different numbers of tasks and raters were used. Here, the D studies were found to point to confidence in scores even in one-task and one-rater condition. To investigate possible problems with this finding, average exact (where the raters gave the same score), adjacent (where the rater scores differed by 1 point) and nonadjacent (where the rater scores differed by 2 or more points) agreement rates between the four raters were calculated, and individual test taker scores were analyzed through cross-classifications of scores assigned by different raters.

Results

Scoring validity is discussed in relation to the parameters proposed by Taylor and Galaczi (2011), namely assessment criteria or the rating scale, the rating process, the rating conditions, the rater characteristics, and the rater training. In what follows a rationale for and the discussion of these parameters are provided. And then for an investigation into scoring validity, G study, D study, rater agreement and cross-classification analyses will be presented.

The Reliability of the Rating Process

Assessment Criteria and Rating Scale

When an assessment involves judgment, the construct is defined by the criteria used to evaluate performance (Brown, 2005, cited in Ducasse & Brown, 2009). Thus, the way rating scales and rating criteria are interpreted may "act as *de facto* test constructs" (McNamara, Hill, & May, 2002, p. 229) and the rating scale provides a test developers' conceptualization of the construct being assessed. Since the development of the test tasks in the current study was informed by the CEFR, the level descriptors were also selected from the CEFR. As Weir (2005b) points out, the CEFR level descriptions are "the least arbitrary sequence of scaled proficiency descriptors available to us at the moment" (p. 282).

An important decision that had to be made in the development of the rating scale was whether to develop different scales for each task or use a generic scale for all of the six tasks. Fulcher (1996) argues that most score variance can be explained by test taker ability when rating scales do not refer to specific tasks. Similarly, Fulcher and Reiter (2003) note that when the rating scale is not task specific, ability contributes more to score variance than task conditions. Therefore, the final decision was to use a generic scale that makes no reference to task conditions. Any level description in the CEFR that could refer to task types, task conditions or tasks were excluded from the scale. Being based on the CEFR, the rating scale was an analytic one which had the advantages of providing specific information about the strengths and weaknesses of test takers as well as enhancing rater agreement and rater reliability (Luoma, 2004; Weir, 1993).

Rating Process and Rating Conditions

The rating process is about the decision-making processes that raters go through (Taylor & Galaczi, 2011). In order to simplify this process for the raters, the scale descriptors should be positive, definite, clear, independent and brief (Pollitt & Murray, 1996). Since the rating scale was adapted from the CEFR scale descriptors, it was assumed that the scale descriptors would provide clear and definite descriptions, helping ease the scoring process for the raters. However, rater perceptions of the descriptors were not explicitly investigated in the current study.

In terms of rating conditions, Taylor and Galaczi (2011) emphasize the temporal, spatial, and psychological dimensions of different rating conditions. In the current study, the raters did not have to make real-time judgments; instead, they had the opportunity to listen to the performances several times on the audio recordings before arriving at a decision about test taker performances. Therefore, they had relatively little cognitive load in comparison to a testing condition where raters have to make real-time judgments. It is likely that having the opportunity to replay the whole or parts of the audio recording of the test takers allowed the raters to make finer distinctions between the performances, which must have contributed to the high rater agreement that will be discussed under the second research question.

Rater Characteristics and Rater Training

Possible rater effects that may threaten the scoring validity of a test are: (a) rater leniency or severity, (b) the halo effect (for example, assigning all analytic scores based on an overall impression), (c) inconsistency in applying the analytic dimensions, and (d) rater bias against a sub-population of test takers (Myford & Wolfe, 2004). These aspects of the rater characteristics were addressed through rater training and rater monitoring. The rater training session on the rating scale and scoring procedures, which lasted a total of six hours, took place before the actual scoring. Test taker performances from the previous pilot tests were used to exemplify performances at different levels for each dimension in the rating scale. Following the ratings, the scores were compared by the researcher and in case of discrepancies the raters were asked to rate for a second time the particular performances on which there were disagreements. The use of separate analytic scores on the four scoring dimensions seems to have contributed to agreement among the raters.

Score Dependability

Table 2 shows the descriptive statistics for the analytic scores of the test takers by task. It should be noted that each task was taken by a total of 12 learners. The means of different dimensions and of different task and dimension combinations were quite close. In tasks 4 and 5, the mean scores were relatively lower than the means of the other tasks; however, the difference was not substantial. Overall, the means of the tasks in Set B (Task 4, 5, 6) were lower than those in Set A (Task 1, 2, 3).

Table 2. Descriptive statistics for the analytic scores.

	Dimension	M	SD
Task 1	F	7.47	3.60
	A	7.37	3.19
	L	7.27	3.30
	C	7.91	3.16
	Total	7.51	3.40
Task 2	F	7.41	3.53
	A	7.08	3.37
	L	7.08	3.28
	C	7.77	3.34
	Total	7.33	3.46
Task 3	F	7.35	3.57
	A	7.54	3.15
	L	7.29	3.26
	I	8.10	3.35
	Total	7.57	3.42
Task 4	F	6.52	2.22
	A	6.77	2.31
	L	6.81	2.13
	C	7.54	2.27
	Total	6.91	2.25
Task 5	F	6.58	2.51
	A	6.47	2.18
	L	6.33	2.14
	C	7.16	2.31
	Total	6.64	2.34
Task 6	F	7.00	2.55
	A	7.14	2.65
	L	7.06	2.49
	I	7.72	2.60
	Total	7.23	2.65

Note. F = fluency, A= grammatical range and accuracy, L = lexical resource, C = coherence, I = interaction; M = mean, SD = standard deviation. Minimum Score Assigned= 1, Maximum Score Assigned = 12; Maximum Possible Score = 12.

Univariate G Analyses on the Dependability of Analytic Scores

For the G analysis, Coherence and Interaction scores were pooled into the Coherence/Interaction (C/I) score since the test takers were scored on Coherence in the first two tasks in both sets, and Coherence was replaced with Interaction in the third tasks. Based on a fully crossed p x t x r design for the four scoring dimensions, Table 3 shows the variance components and the percentages of the total variance accounted for by each source of variance. The analyses were carried out separately for each scoring dimension to investigate score dependability for each of the analytical dimensions.

Table 3. Variance components for the p x t x r design.

Sources of variation		Variance component				Percent of total variation			
		F	A	L	C/I	F (%)	A (%)	L (%)	C/I (%)
SET A	P	13.383	10.970	11.351	11.60	96.	95.1	95.9	97.5
	T	.005	.047	.013	7	0.0	0.4	0.1	0.1
	R	.149	.101	.101	.011	1.1	0.9	0.9	0.1
					.006				
	PT	.056	.007	.005	.046	0.4	0.1	0.0	0.4
	PR	.128	.202	.166	.086	0.9	1.8	1.4	0.7
	TR	.007	.004	.004	.014	0.1	0.0	0.0	0.1
	PTR	.163	.196	.198	.129	1.2	1.7	1.7	1.1
SET B	P	6.204	5.936	5.284	6.011	94.7	93.7	92.4	95.2
	T	.055	.097	.118	.063	0.9	1.5	2.1	1.0
	R	.011	.001	.000	.037	0.0	0.0	0.0	0.6
		.083	.148	.063	.042	1.3	2.3	1.1	0.7
	PT	.044	.005	.013	.000	0.7	0.1	0.0	0.0
	PR	.018	.002	.033	.027	0.3	0.0	0.6	0.4
	TR								
	PTR	.147	.146	.220	.129	2.3	2.3	3.9	2.1

Note. P = person, T = task, R = rater, PT = person-by-task, PR = person-by-rater, TR = task by rater, PTR = person-by-task by rater; F = fluency, A = grammatical range and accuracy, L = lexical resource, C/I = coherence/interaction.

In Table 3, the F, A, L and C/I columns under ‘Variance component’ present variance component values associated with sources of variance (persons, tasks and raters) and their interactions (person-by-task interaction, person-by-rater interaction, task-by-rater interaction, and person-by-rater-by-task interaction plus undifferentiated error) for fluency (F), grammatical range and accuracy (A), lexical resource (L), and coherence/interaction (C/I) dimensions, respectively. The ‘Percent of total variation’ columns show these variance component values converted to a percentile scale for ease of interpretation. These columns, therefore, present the proportion of variance of individual scores that is attributable to each variance source. As explained by Xi and Mollaun (2006), Persons (or test takers) constitute the object of measurement, not error, and indicate systematic individual differences in test taker ability. The other sources of variation are considered sources of error. The task main effect (T) indicates differences in difficulty levels of the tasks; so, if the proportion of variance attributed to tasks is large, that means the test takers’ scores differed across the tasks. The rater main effect (R) indicates raters’ leniency or harshness in their ratings, that is the extent to which the mean scores assigned by different raters to the same performance are the same. Large rater main effect can be interpreted as differences in rater judgments. Interaction effects, on the other hand, indicate (in)consistency in the rank ordering of test takers. A large person-by-rater interaction, for example, indicates that the test takers are rank-ordered differently by different raters.

As can be seen in Table 3, the test takers showed similar performances in their fluency, accuracy, lexical range and coherence/interaction in both sets, as indicated by the similar variance components associated with persons (true variance in CTT) on these four dimensions. That is, the variance associated with the persons (test takers) was similar in all of the four dimensions. Overall, a substantial proportion of the total variance in scores on the four dimensions could be explained by real differences in the test takers' fluency, accuracy, lexical range, or coherence/interaction. Among the four dimensions, 97.5% in Set A and 95.2% in Set B of the variance in the test takers' coherence/interaction scores was explained by variance associated with persons, suggesting that the test takers' scores in this dimension were the most reliable; yet, it should be noted that the percent of variation explained by real differences in person ability was over 90% in all task-by-dimension combinations.

These results suggest that the test takers were rank ordered similarly on each of the four dimensions by the four raters; so, the raters did not differ much in judging where a test taker stood compared to the other test takers. In other words, in only a few cases, the test takers were rank ordered differently by the raters. The small rater main effects indicate that the raters did not differ much in their leniency or harshness. The task main effect and the person-by-task interaction were almost zero, indicating that the mean scores of this group of test takers were similar across the tasks (i.e., on average, the tasks varied little in their difficulty levels) and the scores assigned to the test takers were similar on each of the four dimensions across the tasks. The task-by-rater interaction was negligible too, pointing to minimal difference in the raters' rank orderings of task difficulty. This finding may be attributed to a number of factors. Firstly, the G coefficient will be higher when the sample size is small and when the target population is more heterogeneous (Xi & Mollaun, 2006). The number of test takers in the two sets was not very large (12 test takers took each set), so the sample size was small. In addition, the test takers in both sets were from a range of different proficiency levels and hence formed a heterogeneous group in terms of proficiency. Moreover, the scores were assigned following rater monitoring, which attempted to minimize discrepancies among the raters.

Overall, the G analyses indicated that the raters did not differ much in their leniency or harshness. This may be attributed to rater training and monitoring or to small sample size. The tasks were expected to be of different difficulty levels; however, the G analyses suggested the opposite, that is, the tasks were found to be of similar difficulty. In other words, a particular task taker performed similarly and was rated with similar scores in different tasks.

D Studies: Changes in Phi Coefficients

It should be noted that in the D study, "the values determined for the estimated variance components [in the G study] are used to further calculate estimates of the effects of various measurement designs on the *dependability* (analogous to *reliability*) of the scores" (Brown, 2005, p. 13). In the current study, D studies were conducted where the number of tasks and raters were varied to examine their impact on the phi coefficients[5] of the analytic scores (Xi & Mollaun, 2006) so that the number of tasks and raters could be determined for more dependable results.

Table 4 provides the phi coefficients for the analytic scores when different combinations of number of tasks and raters are used for a fully crossed p x T x R design for both Set A and Set B.

Table 4. Changes in phi coefficients of the four analytic scores in D studies.

Alternative D studies for p x T x R design									
		Single rating				Double rating			
	# of tasks	F	A	L	C/I	F	A	L	C/I
SET A	1	.963	.951	.958	.975	.979	.972	.978	.985
	2	.971	.962	.967	.983	.984	.979	.983	.990
	3	.974	.965	.970	.986	.986	.981	.984	.992
	4	.975	.967	.972	.987	.987	.982	.985	.993
	5	.976	.968	.973	.988	.987	.983	.986	.993
	6	.976	.969	.973	.989	.988	.984	.986	.994
SET B	1	.946	.937	.923	.952	.962	.948	.944	.967
	2	.969	.966	.960	.972	.978	.973	.971	.981
	3	.976	.977	.973	.979	.984	.981	.980	.986
	4	.980	.982	.979	.983	.987	.986	.985	.989
	5	.983	.985	.983	.985	.989	.988	.988	.990
	6	.984	.987	.986	.986	.990	.990	.990	.991

Note. F = fluency, A = grammatical range and accuracy, L = lexical resource, C/I = coherence/interaction.

In Table 4, the F, A, L, and C/I columns under 'single rating' show what the phi coefficient values would be if test taker performances were rated only once with different numbers of tasks used (from 1 to 6 tasks). For example, for Fluency, with one rating and one task, the obtained phi coefficient would be .963, while it would increase to .976 with one rating and six tasks. This indicates that using six tasks produces more dependable scores. However, .963 already indicates high dependability. The columns under 'double rating' show how the phi coefficients would change if the performances were rated twice with different numbers of tasks. As can be seen from Table 4, the phi coefficient is substantial even when one task and one rater are used since, as was observed in G study results, the tasks did not differ much in their difficulty levels, the test takers obtained similar scores across the tasks, and the raters awarded consistent scores across the test takers and tasks.

[5] The g- and phi-coefficients reported in generalizability analyses can both be interpreted as reliability coefficients. They describe how well the mean rating for an individual examinee predicts her universe score (Sudweeks, Reeve and Bradshaw, 2005, p. 244).

One obvious observation from D studies was that the phi coefficients increase when more raters and more tasks are used. When one rating is obtained for each task, using two or three tasks yields higher phi coefficients than with one task, but as the number of tasks increases from four to six, the improvement in phi coefficients is less dramatic.

Dependability of Analytic Scores by Task

The analyses above illustrate how the phi coefficients would change with different combinations of number of tasks and raters. The analyses were based on averaged variance components for all three tasks in each set. Analysis of the dependability of analytic scores at the task level was also carried out in order to examine which tasks introduced more unreliability in scoring a particular dimension (Xi & Mollaun, 2006). At this stage, each task was analyzed separately for variance components for the four scoring dimensions. Each analysis featured a fully crossed p x r design. The variance components associated with different sources of variance for different tasks are shown in Table 5.

Table 5. G study variance components by tasks.

		Variance component				Percent of total variation			
		F	A	L	C/I	F	A	L	C/I
Task 1	P	13.703	10.713	11.393	10.766	96.4%	96.0%	95.9%	98.4%
	R	.253	.093	.093	.010	1.8%	0.8%	0.8%	0.0%
	PR	.253	.351	.399	.176	1.8%	3.1%	3.4%	1.6%
Task 2	P	13.361	11.690	11.217	11.902	98.1%	94.0%	95.4%	97.3%
	R	.002	.179	.122	.053	0.0%	1.4%	1.0%	0.4%
	PR	.252	.570	.419	.273	1.9%	4.6%	3.6%	2.2%
Task 3	P	13.417	10.531	11.327	12.056	96.3%	97.1%	97.4%	98.2%
	R	.171	.044	.075	.012	1.2%	0.4%	0.7%	0.1%
	PR	.349	.275	.229	.202	2.5%	2.5%	2.0%	1.7%
Task 4	P	5.118	5.685	4.573	5.444	94.9%	97.6%	92.0%	96.8%
	R	.002	.010	.046	.011	0.0%	0.0%	0.9%	0.2%
	PR	.273	.142	.349	.169	5.1%	2.4%	7.0%	3.0%
Task 5	P	6.782	5.102	4.864	5.577	98.4%	97.7%	97.0%	95.3%
	R	.000	.006	.051	.161	0.0%	0.1%	1.0%	2.8%
	PR	.111	.111	.101	.116	1.6%	2.1%	2.0%	2.0%
Task 6	P	6.936	7.465	6.606	7.291	97.3%	97.4%	97.4%	98.6%
	R	.032	.001	.001	.022	0.5%	0.0%	0.0%	0.3%
	PR	.161	.202	.172	.081	2.3%	2.6%	2.5%	1.1%

Note. P = person, R = rater, PR = person-by-rater; F = fluency, A = grammatical range and accuracy, L = lexical resource, C/I = coherence/interaction.

Overall, the variances explained by sources of error were very small. The task with the largest source of error associated with person-by-rater interaction was Task 4, followed by Task 2, which are both picture description tasks. Task 4 required the test takers to describe different types of holidays shown in the pictures and then choose one type of holiday, explaining why they would prefer that particular holiday type over the others. Task 2 required the test takers to describe two events shown in the pictures and then to make predictions as to what might have happened before the picture was taken and what may happen afterwards. It needs to be noted again that the error variance in all of the tasks was very small, including these two picture description tasks.

Table 6 shows the results of the D studies, where phi coefficients were compared for single versus double raters for a p x R design. Substantially high phi coefficients were obtained in all task-dimension combinations and the phi coefficients were very close in all combinations. Results clearly display that the phi coefficient increases when two raters are used.

Table 6. Changes in phi coefficients by task.

		Single rating				Double rating			
		F	A	L	C/I	F	A	L	C/I
Phi coefficient	Task 1	.981	.960	.958	.983	.990	.979	.978	.991
	Task 2	.981	.939	.953	.973	.990	.968	.976	.986
	Task 3	.962	.970	.973	.982	.980	.985	.986	.991
	Task 4	.949	.975	.920	.967	.973	.987	.958	.983
	Task 5	.983	.977	.969	.952	.991	.988	.984	.975
	Task 6	.972	.973	.974	.985	.986	.986	.987	.992

Note. F = fluency, A = grammatical range and accuracy, L = lexical resource, C/I = coherence/interaction.

Rater Agreement

In this section, the scores assigned by the raters are examined in more detail. Table 7 shows the rater agreement rate by dimension and by task.

Table 7. Average agreement rates between the four raters on each scoring dimension by task

Task	Fluency			Gram. Range & Acc.			Lexical Resource			Coherence/Interaction		
	Ex. %	Adj. %	Non. %	Ex. %	Adj. %	Non. %	Ex. %	Adj. %	Non. %	Ex. %	Adj. %	Non. %
1	44.4	40.3	15.3	51.4	37.5	11.1	56.9	31.9	11.1	66.7	33.3	0.0
2	50.0	50.0	0.0	43.1	36.1	20.8	48.6	34.7	16.7	50.0	44.4	5.6
3	33.3	50.0	16.7	63.9	27.8	8.3	65.3	26.4	8.3	52.8	47.2	0.0
4	54.2	43.1	2.8	72.2	27.8	0.0	41.7	48.6	9.7	63.9	36.1	0.0
5	77.8	22.2	0.0	76.4	23.6	0.0	69.4	30.6	0.0	52.8	44.4	2.8
6	61.1	38.9	0.0	59.7	40.3	0.0	65.3	34.7	0.0	79.2	20.8	0.0
Total	53.5	40.7	5.8	61.1	32.2	6.7	57.9	34.5	7.6	60.9	37.7	1.4

Note. Ex. = Exact, Adj. = adjacent, Non. = nonadjacent.

The rater agreement rates were higher for the tasks in Set B (Task 4, 5 and 6). The combined agreement rates were similar for all of the four dimensions. We also examined nonadjacent discrepancy, where the rater scores differed by 2 or more points. The overall nonadjacent discrepancy rate for coherence/interaction was the lowest among the four dimensions, indicating that the raters had the largest agreement rate for this dimension. The nonadjacent agreement rates for Lexical Range, Grammatical Range and Accuracy, and Fluency were slightly higher. Some of the largest percentages of nonadjacent discrepancies occurred with Fluency in Task 1 and Task 3, Grammatical Range and Accuracy and Lexical Resource in Task 2. Most of these nonadjacent discrepancies were mainly associated with two rater pairs, Rater 2 and Rater 3, and Rater 3 and Rater 4, which accounted for 30 of 50 discrepancies. This raises questions about including only one rater in the assessment process, even though the D studies produced high dependability coefficients for one rater. Moreover, it should be noted that these results were obtained following rater monitoring where the aim was to remove the discrepancies by asking the raters rate for a second time the performances for which discrepancies were observed.

Cross-Classifications of the Ratings

All of the analyses so far have provided summary statistics for raters, tasks, criteria or the interactions between them. Despite important discrepancies between two rater pairs, overall high agreement was obtained between the four raters. The G and D studies also produced high generalizability and dependability coefficients, pointing to confidence in scores even with a one-rater condition. However, a closer investigation into rater behavior points to problems with this interpretation, indicating that even very high indices of inter-rater correlation or high dependability coefficients may not justify using one rater only. Cross-classifications of scores assigned to the test takers by different raters indicated a few important discrepancies between the rater scores. Table 8 and Table 9 present the cross-classification of rating frequencies for two different pairs of raters based on the scores they awarded to two different test takers.

Table 8. Cross-classification of rating frequencies for raters 2 and 3 for Anna.

	Rater 3 (score assigned)					
Rater 2 (score assigned)	7	8	9	10	11	Row total
7						0
8						0
9			1			1
10		4	3			7
11	1	1	1	1		4
Column Total	0	5	5	1	1	12

Table 9. Cross-classification of rating frequencies for raters 1 and 2 for Jonathan.

	Rater 2 (score assigned)				
Rater 1 (score assigned)	2	3	4	5	Row total
2					0
3					0
4	1	4	2		7
5	1	1	3		5
Column Total	2	5	5	0	12

Raters 2 and 3 agreed on 1 out of 12 occasions, and substantial discrepancies were observed on certain dimensions. For example, while Rater 2 assigned 11 on Grammatical Range and Accuracy to Anna (a pseudonym) in Task 2, Rater 3 provided a score of 7. In 7 out of 12 ratings, the two raters differed by 2 or more points. Similar findings were observed for a few of the other test takers as well. For example, Raters 1 and 2 differed substantially in terms of the scores they assigned to Jonathan (a pseudonym). They agreed on 2 out of 12 cases, and a few of the discrepancies were especially noteworthy. For example, while Rater 1 gave a score of 5 to Jonathan on Lexical Range in Task 2, the rating given by Rater 2 on the same criteria was only 2. This finding suggests that at least two raters need to be employed for ensuring score dependability for each individual test-taker.

Discussion

The generalizability analysis indicated that the analytic scores would be sufficiently reliable for both low-stakes practice settings and operational use. The D-studies indicated that the phi coefficients of the analytic scores were fairly high even for one task with a single rating (.963 – .976) and they improved with double ratings. The improvement was leveled off with each additional task and rater, which indicated that the test may not need to include six tasks or more than one rater to produce dependable results.

G studies on the scores showed that discrepancies in rater leniency (or severity) accounted for a negligible amount of error (it explained 1.1% of the variance in Fluency scores in Set A at its highest). This may be attributed to rater monitoring, because the raters were asked to re-rate the performances for which there were substantial discrepancies in their first ratings. The raters were also consistent in their rank ordering of examinees, the person-by-rater interaction explaining 0.9%, 1.8%, 1.4% and 0.7% (in Set A) and 0.7%, 0.1%, 0.0% and 0.0% (in Set B) of the variance of fluency, grammatical range and accuracy, lexical resource, and coherence/interaction scores, respectively. These results indicate that the raters showed very few discrepancies in the absolute scores they assigned to the test takers.

The variance associated with the task main effect and person-by-task interaction was also very low. Although it has been noted in the language assessment literature that some tasks may pose more difficulty for particular sub-groups of test takers (Dunbar, Koretz, & Hoover, 1991), varying results have been reported as to the issue of what makes a task more difficult for a certain group of test takers and whether differences in task characteristics are reflected in test scores (Bachman, Lynch & Mason, 1995; Elder et al., 2002; Fulcher, 1996b). In the current test, although the tasks were expected to be of different levels of difficulty, this was not reflected in the scores. Lack of score sensitivity to variations in task conditions has been noted in other research studies as well (Fulcher, 1996b; Fulcher & Reiter, 2003; Bachman, Lynch, & Mason, 1995; Lumley & O'Sullivan, 2005). In other words, similar task difficulty levels were observed for different types of tasks such as role play and individual long turn (Bachman, Lynch, & Mason, 1995), interview and group discussion (Fulcher, 1996b). Moreover, variations in task conditions such as social power and degree of imposition (Fulcher & Reiter, 2003) or task familiarity (Lumley & O'Sullivan, 2005) did not produce significantly different scores for individual test takers. In addition, Iwashita, McNamara and Elder (2001) found that differences in assumed degree of cognitive demand based on Skehan's (1998) framework did not predict discourse or score variation. Reviewing a number of studies that have investigated task difficulty, Fulcher and Reiter (2003) write, "research has consistently shown that it requires gross changes in task type to generate significant differences in difficulty from one task to another, and even then the task accounts for little score variance" (p. 326). The findings of the present study also suggest that variations in tasks may not translate into changes in task scores. Similarly, Xi and Mollaun (2006) argue that variation in performance across tasks is less likely if "an assessment uses tasks that are not very differentiated in task types and in the ways tasks are contextualized and uses scoring criteria that are more driven by components that are relatively stable across tasks". (p. 37-38). Thus, they point out that very contextualized tasks and task-specific scoring criteria are likely to result in large person-by-task interaction.

The little task main effect and person-by-task interaction found in the current analysis may be attributed to various reasons. First of all, it may be because of the scoring criteria used. The same scoring criteria were used to rate performances on all tasks (except, of course, that the Coherence dimension was replaced with Interaction in the two tasks that require the test takers to interact with each other). The rating scale does not contain any features that are task-specific; on the contrary, effort was invested in excluding any criteria that may apply to one task but not another. Fulcher (1996b) proposes that when rating scales do not make reference to specific task types, task conditions or tasks, learner ability can account for most variance within the scores assigned. A second possibility is that the tasks were of similar difficulty for this particular group of test takers who participated in the current test, but future administrations of the tasks to different groups may produce more person-by-task variation. A further reason for the little variance attributed to the task main effect and person-by-task interaction may be that the raters might have tended to assign similar scores to the test takers across tasks based on their overall impression about them on a particular task or the overall test. Teng (2007), for example, found that different scoring methods could show different levels of sensitivity to different task types. She found no difference among different task scores with holistic scoring, while analytic scores produced significant task main effects for complexity and fluency, but not for accuracy. In the present study, in order to minimize possible halo effects, the raters were asked to rate the performances of all test takers on the first task and then continue with the second task and the third task, instead of rating one test taker on all three tasks and continuing with the next test taker. However, if the raters still assigned similar scores across different tasks based on their holistic judgments of test taker performances, this may be one of the reasons for the absence of task main effects.

Overall, the analyses showed that test the taker performances were similar across tasks based on the scoring criteria used in the test. It should also be noted here that these findings are consistent with current theoretical models of communicative competence, which, while recognizing the possible influence of context and local components, assert that communicative competence is stable to some extent (Chalhoub-Deville, 2003).

It has been pointed out earlier that G theory addresses the statistical question of the consistency of test taker performance across various tasks and task types. It provides information as to whether the scores obtained from a combination of different tasks can be generalized to the universe of tasks which are similar to those included in the assessment. However, Xi and Mollaun (2006) note:

> G theory can by no means provide evidence for establishing the link between performance on a sample of tasks (“observed score”) and expected performance in the target domain (“target score”), unless there is ample evidence to support that the universe of generalization and the target domain are similar. (p. 39)

From the beginning of the test development process, the issue of validity was the overarching consideration in the current test. Still, another important implication to be drawn from Xi and Mollaun’s argument is that although a high phi coefficient in the D studies was obtained for one task and one rater condition, using only one task could not be justified because one task would by no means provide enough evidence to argue that the universe of generalization and the target domain are similar. Therefore, despite the statistical results obtained, several tasks need to be used in speaking tests to ensure generalizability.

With regards to the raters, the G studies revealed small rater main effects, person-by-rater interactions and task-by-rater interactions. However, because of the small sample size (12 test takers in each set of the test), which influences results obtained from G and D studies, interpretations based only on these statistics need to be made cautiously. For example, the investigation of the average agreement rates between the four raters on dimension-by-task combinations revealed some discrepancies between the raters. Overall, the nonadjacent agreement rates were relatively high and the discrepancies between ratings were more substantial for a few of the dimension-by-task combinations in comparison to others. These discrepancies, however, were produced mainly by two rater pairs out of six rater-rater pairings. In general, the results obtained from the analysis of agreement indicated large agreement rates among the four raters. To better understand the extent of the discrepancies observed in the agreement rate analysis, the scores assigned to each individual test taker were separately analyzed. Cross-classification of rating frequencies revealed using one rater could not be justified despite the very high inter-rater correlations, large agreement rates among the raters, or the D study results which produced high phi coefficients for the one task-one rater condition. This is because the summary statistics did not reveal differences at individual level, which was indicated by discrepancies in the scores assigned to specific test takers on particular dimensions. Therefore, these analyses suggested that at least two raters and a number of different tasks need to be used for fairness and validity considerations.

Conclusion

The scoring validity of a speaking test in TSL was analyzed based on the parameters proposed by Taylor and Galaczi (2011) and using Generalizability Theory. The descriptions of the levels in the rating scale were adapted from the CEFR level descriptions. The issue of rater variability was addressed through rater training and monitoring. Still, a few discrepancies were noted between the scores assigned by the raters. To analyze rater effects, the dependability of analytic scores and the effect of the number of tasks and raters on score dependability, generalizability and decision studies were carried out. It is important to note that given the small data set, the conclusions drawn are tentative. First of all, high score dependability was obtained for all of the six tasks used. The D studies indicated high score dependability even for one task and one rating condition. However, problems with this interpretation were noted. The one-task condition could not be justified, partly because each task requires different language resources, discourse and interaction types. Moreover, despite the very high inter-rater correlations, an investigation into the scores assigned to individual test takers by different raters through cross-classification of the ratings showed that there were important discrepancies in a few cases, and this suggested that at least two raters need to be used for fairness and validity considerations. This concern can be justified also because these discrepancies were observed after rater training and rater monitoring. Besides, the analyses showed that the scoring dimensions did not produce different scores from each other, but they were not in complete agreement, either. It was argued that this finding could be attributed to the sample of test takers, the nature of the analytic rubric used, the quality of analytic ratings, or a combined effect of these.

One of the limitations of the current study is the limited sample size. The conclusions drawn from the analyses are, therefore, tentative. In addition, the rating scale was adapted from the CEFR level descriptors, but with a bigger performance data, the scale could be based on empirical findings from the L2 Turkish performance of test takers with varying proficiency levels. This in an area in which future research is necessary. Still, the rating scale developed within the current study based on the CEFR level descriptors was shown to help produce reliable scores with the substantial care taken for the aspects of the testing that may impact upon the rating process. The G and D analyses have shown that the scores obtained were dependable, which substantiates the scoring validity arguments for the test. To the best of our knowledge, this is the first validation study for a rating scale for TSL. Thus, it offers a starting point for the development of such a rating scale using and revising it in new testing conditions with new tasks and raters and a larger number of test takers.

References

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

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

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

Bachman, L. F., & Savignon, S. J. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL oral interview. *The Modern Language Journal*, *70*(4), 380-390.

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing*, *12*(2), 238-257.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing*, *12*(2), 86-107.

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement*, *24*(4), 339-353.

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

Brown, J. D. (1999). The relative importance of persons, items, subtests and languages to TOEFL test variance. *Language Testing*, *16*(2), 217-238.

Brown, J. D. (2005). Statistics corner, questions and answers about language testing statistics: Generalizability and decision studies. *Shiken: JALT Testing & Evaluation SIG Newsletter, 9*(1), 12–16. Retrieved from http://jalt.org/test/bro_21.htm

Brown, J. D., & Ahn, R. C. (2011). Variables that affect the dependability of L2 pragmatics tests. *Journal of Pragmatics*, *43*(1), 198-217.

Brown, J. D., & Kondo-Brown, K. (2012). Rubric-based scoring of Japanese essays: The effects on generalizability of numbers of raters and categories. In J. D. Brown (Eds.), *Developing, using, and analyzing rubrics in language assessment with case studies in Asian and Pacific languages* (pp. 169-182). National Foreign Language Resource Center: University of Hawai'i at Manoa.

Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing, 28*(2), 201–219.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing*, *20*(4), 369-383.

Chalhoub-Deville, M. (2006). Drawing the line: The generalizability and limitations of research in applied linguistics. In M. Chalhoub-Deville, C. A. Chapelle, & P. Duff (Eds.), *Inference and generalizability in applied linguistics: Multiple perspectives* (pp. 1-5). Amsterdam: John Benjamin Publishing Company.

Chalhoub-Deville, M., & Wigglesworth, G. (2005). Rater judgment and English language speaking proficiency. *World Englishes*, *24*(3), 383-391.

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

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing*. Cambridge: Cambridge University Press.

Deville, C., & Chalhoub-Deville, M. (2006). Old and new thoughts on test score variability: Implications for reliability and avlidity. In M. Chalhoub-Deville, C. A. Chapelle, & P. Duff (Eds.), *Inference and generalizability in applied linguistics. Multiple perspectives*. (pp. 9- 25). Amsterdam: John Benjamin Publishing Company.

Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing*, *26*(3), 423-443.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education*, *4*(4), 289-303.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, 2, 197–221.

Eckes, T. (2009). Many-facet Rasch measurement. *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*. Retrieved from http://www.coe.int/t/dg4/Linguistic/CEF-refSupp-SectionH.pdf

Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Frankfurt, Germany: Peter Lang.

Educational Testing Service (2004). *IBT/ Next generation TOEFL test independent speaking rubrics (scoring standards)*. Retrieved from http://www.ets.org/Media/Tests/TOEFL/pdf/Speaking_Rubrics.pdf

Elder, C., Iwashita, N. & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: What does the test-taker have to offer? *Language Testing, 19*(4), 337-346.

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly*, *2*(3), 175-196.

Field, J. (2011). Cognitive validity. In L. Taylor (Ed.), *Examining Speaking: Research and practice in assessing second language speaking* (pp. 65-111), Cambridge: Cambridge University Press.

Fulcher, G. (1996a). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing, 13*(2), 208-238.

Fulcher, G. (1996b). Testing tasks: Issues in task design and the group oral. *Language Testing*, *13*, 23–51.

Fulcher, G. (2003). *Testing second language speaking,* Harlow: Longman/Pearson Education Ltd.

Fulcher, G., & Reiter, R. M. (2003). Task difficulty in speaking tests. *Language Testing, 20*(3), 321-344.

Galaczi, E., & ffrench, A. (2011). Context validity. In L. Taylor (Eds.), *Examining Speaking: Research and practice in assessing second language speaking* (pp. 112-170), Cambridge: Cambridge University Press.

Gülle, T. (2015). *Development of a speaking test for second language learners of Turkish.* (Unpublished Master's Thesis). Boğaziçi University, Istanbul, Turkey.

Hasselgreen, A. (2004). *Testing the spoken English of young Norwegians: A study of testing validity and the role of smallwords in contributing to pupils' fluency*. Cambridge: Cambridge University Press.

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods, 5*, 64–86.

Huang, J. (2012). Using generalizability theory to examine the accuracy and validity of large-scale ESL writing assessment. *Assessing Writing*, *17*(3), 123-139.

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

International English Language Testing System. (2009). *SPEAKING: Band descriptors (public version),* retrieved from https://www.ielts.org/pdf/SpeakingBanddescriptors.pdf

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, *10*(2), 135-159.

Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to test design. *Language Learning, 51*(3), 401-436.

Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing*, *26*(4), 485-505.

Khalifa, H., & Weir, C. J. (2009). *Examining Reading: Research and Practice in Assessing Second Language Reading*, Studies in Language Testing 26, Cambridge: Cambridge University Press.

Kim, Y. H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing*, *26*(2), 187-217.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, *19*(1), 3-31.

Lee, Y. W. (2005). *Dependability of scores for a new ESL speaking test: Evaluating prototype tasks*. TOEFL® Monograph MS-28. Princeton, NJ: ETS.

Lee, Y. W. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing*, *23*(2), 131-166.

Lee, Y. W., & Kantor, R. (2005). Dependability of new ESL writing test scores: Evaluating prototype tasks and alternative rating schemes. *ETS Research Report Series*, *2005*(1), i-76.

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

Lumley, T., & O'Sullivan, B. (2005). The effect of test-taker gender, audience and topic on task performance in tape-mediated assessment of speaking. *Language Testing*, *22*(4), 415-437.

Luoma, S. (2004) *Assessing speaking.* Cambridge: Cambridge University Press.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, *15*(2), 158-180.

McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.

McNamara, T., Hill, K., & May, L. (2002). Discourse and assessment. *Annual Review of Applied Linguistics*, *22*, 221-242.

McNamara, T.F. & Adams, R.J. (1991). Exploring rater behavior with Rasch Techniques. Paper presented at the Annual Language Testing Research Colloquium (Princeton, NJ, March 21–23). Eric document ED345 498.

Messick, S. (1989) Validity. In R. L Linn (Eds.), *Educational measurement* (pp. 13-103). New York: Macmillan/American Council on Education.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*(4), 386-422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*(2), 189-227.

Politt, A., & Murray, N. L. (1996). What do raters really pay attention to? In M. Milanovic, & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium* (pp. 74–91). Cambridge: Cambridge University Press.

Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite. *Language Testing* 24(3), 355–90.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, *44*(6), 922.

Shaw, S. D. & Weir, C. J. (2007). *Examining Writing: Research and Practice in Assessing Second Language Writing.* Studies in Language Testing 26, Cambridge: UCLES/Cambridge University Press.

Shi, L. (2001). Native-and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing*, *18*(3), 303-325.

Skehan, P. (1998). *A Cognitive Approach to Language Learning.* Oxford: Oxford University Press.

Taylor, L. (2011). *Examining Speaking: Research and practice in assessing second language speaking*, Studies in Language Testing 30, Cambridge: Cambridge University Press.

Taylor, L., & Galaczi, E. (2011). Scoring validity. In L. Taylor (Eds.), *Examining Speaking: Research and practice in assessing second language speaking* (pp. 171-233), Studies in Language Testing 30, Cambridge: Cambridge University Press.

TELC (2013). *Diller İçin Avrupa Ortak Öneriler Çerçevesi Öğrenim, Öğretim ve Değerlendirme*. Frankfurt: Telc GmbH.

Teng, H. (2007). *A study of task types for L2 speaking assessment*. (ERIC Document Reproduction Service No. ED496075). Retrieved from http://eric.ed.gov/?id=ED496075

University of Cambridge ESOL Examinations. (2015). *Information for candidates*. Retrived from http://www.cambridgeenglish.org/images/173976-cambridge-english-advanced-examiners-comments.pdf

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, *11*(2), 197-223.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), 263-287.

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

Weir, C. J. (2005a). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.

Weir, C. J. (2005b). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*, *22*(3), 281-300.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, *10*(3), 305-319

Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, *30*(2), 231-252.

Xi, X., & Mollaun, P. (2006). Investigating the utility of analytic scoring for the TOEFL Academic Speaking Test (TAST). *ETS Research Report Series*, i-71.

Xi, X., & Mollaun, P. (2009). How do raters from India perform in scoring the TOEFL iBT™ speaking section and what kind of training helps? *ETS Research Report Series*, *2009*(2), i-37.

Xi, X., & Mollaun, P. (2011). Using raters from India to score a large-scale speaking test. *Language Learning*, *61*(4), 1222-1255.

Zhang, Y. and Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing,* 28(1), 31-50.

Genellenebilirlik Analizi ile İkinci Dil Olarak Türkçe Konuşma Testi Puanlama Geçerliliğinin İncelenmesi

Özet

Test geliştiriciler, test geliştirme ve uygulama süreci boyunca test geçerliğinin tüm boyutlarına önem vermek ve özellikle puanların geçerliğini sağlamak zorundadırlar. Bu çalışma, yabancı bir dil olarak Türkçe'de konuşma becerisini ölçmek için geliştirilen bir testin puanlama geçerliğini incelemiştir. Konuşma becerisini ölçmek üzere altı görev ve bir puanlama ölçeği geliştirilmiş ve Türkçeyi ikinci dil olarak öğrenen yirmi dört öğrenciye uygulanmıştır. Öğrencilerin performansları dört puanlandırıcı tarafindan değerlendirilmiştir. Genellenebilirlik ve Güvenilirlik analizleri ile puan güvenilirliği araştırılmıştır. Sonuçlar, notlardaki varyasyonunun çoğunun test katılımcılarına atfedilebileceğini ve puanlayıcılar veya ödevlerden kaynaklı hata varyansına atfedilemeyeceğini göstermektedir.

Anahtar Terimler: ikinci dil olarak Türkçede konuşma becerisinin ölçülmesi, puanlama geçerliği, genellenebilirlik analizi

Appendix

PUANLANDIRMA ÖLÇÜTÜ

Bu ölçeğin Türkçe çevirisi TELC (2013) tarafından yayınlanan ‘Diller İçin Avrupa Ortak Öneriler Çerçevesi Öğrenim, Öğretim ve Değerlendirme’ metnine dayanmaktadır.

AKICILIK

Puan	CEFR seviyesi	Puan Tanımları
11/12	C2	• Kendini uzun uzun, çaba harcamadan ve duraksamadan, doğal ve akıcı şekilde ifade edebilir. Nadiren tekrar eder veya kendini düzeltir • Yalnızca içeriği kesin bir şekilde ifade etmek için duraksar
9/10	C1	• Kendini pek çaba göstermeksizin anında ve akıcı şekilde ifade edebilir. Yalnıza ara sıra tekrar eder veya kendini düzeltir. • Sadece kavram açısından zor konular dil akıcılığının doğallığını etkileyebilir
7/8	B2	• Anında anlaşabilir, uzun ve karmaşık konuşmalarda da kendini olağanüstü kolaylıkla ve akıcı olarak ifade edebilir • Bazen dille ilgili duraksamalar olabilir, ya da tekrar eder veya kendini düzeltir
5/6	B1	• Anlatımları dilbilgisi ve sözcük bulma açısından planlamak ya da düzeltmek amacıyla belirgin aralar verse de özellikle uzun uzun ve serbest konuştuğunda pek duraksamadan kendini anlaşılır biçimde ifade eder
3/4	A2	• Sık sık duraksamasına, yeniden söze başlamasına ya da başka sözcüklerle tekrarlamasına rağmen, kısa anlatımlarla kendini ifade eder
1/2	A1	• Çok kısa, kalıplaşmış ve çoğunlukla önceden ezberlenmiş anlatımları kullanabilir; ancak sözcük bulmak, az bilinen sözcükleri söyleyebilmek ve bildirişim kopukluklarını gidermek için sık sık ara verir

DİLBİLGİSİ ALANI VE DİLBİLGİSEL DOĞRULUK

Puan	CEFR seviyesi	Puan Tanımları
11/12	C2	• İnce anlam ayrıntılarını belirtmek, bir şeyi vurgulamak, ayrımlaştırmak ya da çokanlamlılıkla başa çıkabilmek için, düşüncelerini yeniden düzenlerken çok esnek bir şekilde çeşitli dilsel araçları kullanabilir • Karmaşık dil kullanırken dilbilgisi hâkimiyetini korur
9/10	C1	• Kapsamlı dil yeterliğine sahip olduğundan seçtiği anlatımlar ile söylemek istediklerini kısıtlamadan kendini açıkça ifade edebilir • Büyük ölçüde dilbilgisi kurallarına sadık kalır; nadir olarak ve pek farkına varılmayan hatalar yapar
7/8	B2	• Yeterli genişlikte dilsel araçlara sahip olduğundan belli etmeden sözcükleri bularak ve birkaç karmaşık cümle yapısını da kullanarak konuşabilir • Dilbilgisi kurallarına iyice hâkimdir; yanlış anlaşılmaya neden olacak hatalar yapmaz
5/6	B1	• Sık kullanılan basmakalıp sözleri ve ifadeleri içeren bir dağarcığı yeterli derecede doğru olarak kullanır ve hatalarının çoğunu düzeltir • Hatalar yapar ancak ne söylemek istediği açıktır
3/4	A2	• Kısa sözcük kümelerini, konuşma kalıplarını ve temel tümce yapılarını ezberlenmiş ifadelerle birlikte kullanır • Bazı basit yapıları doğru olarak kullanır ama yine de sistematik temel yanlışlar yapar
1/2	A1	• Çok kısıtlı düzeyde sözcük dağarcığına ve ifadeye sahiptir • Birkaç basit dilbilgisi yapısını ve kalıp cümleleri kapsayan kısıtlı ezberlenmiş bir dağarcığa sahiptir

SÖZCÜK DAĞARCIĞI

Puan	CEFR seviyesi	Puan Tanımları
11/12	C2	• Çok geniş bir sözcük dağarcığına hâkimdir ve ince anlam farklılıklarını ayırt edebilir • Sözcük dağarcığını sürekli olarak doğru ve uygun bir şekilde kullanır
9/10	C1	• Geniş bir sözcük dağarcığına hâkimdir; çok nadir sözcük arar ya da bilmediği bir şeyi kullanmaktan kaçınır • Sözcük kullanımında bazı küçük pürüzlere rağmen büyük hatalar yapmaz
7/8	B2	• Sık tekrarlamalar yapmamak için, değişik ifadelere başvurabilir; ama buna rağmen sözcük dağarcığındaki eksiklikler duraklamaya ve başka tanımlamalar aramaya yol açabilir • Sözcük dağarcığı genelde doğru olarak kullanılır, bazı karıştırmalar ve yanlış sözcük kullanımları olmasına rağmen bunlar bildirişimi bozmaz
5/6	B1	• Daha karmaşık fikirleri ifade ederken bazı temel yanlışlar yapmasına rağmen, temel sözcük dağarcığına iyice hâkimdir
3/4	A2	• Kısıtlı bir sözcük dağarcığına sahiptir
1/2	A1	• Tek tük sözcük ve deyimlerden oluşan temel bir birikime sahiptir

TUTARLILIK

Puan	CEFR seviyesi	Puan Tanımları
11/12	C2	• Çeşitli bölümleme ve bağlantı kurma olanaklarını uygun bir şekilde kullanarak iyi yapılandırılmış ve bağlantılı bir metin oluşturabilir
9/10	C1	• Anlaşılır, çok akıcı ve iyi yapılandırılmış şekilde konuşabilir ve bölümleme, içerik ve dilsel açıdan bağlantıyı kurabilmek için gerekli olanaklara hâkimdir
7/8	B2	• Anlatımlarını, anlaşılır ve bağlantılı bir metne dönüştürebilmek için az sayıda bağlantı olanakları kullanabilir; ama daha uzun metinlerde kopukluklar oluşabilir
5/6	B1	• Bir dizi kısa ve basit dilsel öğeleri yan yana sıralayarak bağlantılı bir anlatım oluşturabilir
3/4	A2	• Sözcük gruplarını “ve”, “ama”, “çünkü” gibi basit bağlaçlarla birleştirebilir
1/2	A1	• Sözcükleri ve sözcük gruplarını “ve” ya da “sonra” gibi basit bağlaçlarla birleştirebilir

ETKİLEŞİM

Puan	CEFR seviyesi	Puan Tanımları
11/12	C2	• Çok doğal bir şekilde söze girip bir noktaya değinerek, ima ederek vs. kendi görüşlerini konuşma içine katabilir
9/10	C1	• Söze girerek ya da devam ederek veya konuşmalarını başkalarınınkiyle ustaca bağlayarak kendi anlatımını doğru yönlendirmek amacıyla, mevcut söylem araçları dağarcığının içinden uygun bir anlatım biçimi seçebilir
7/8	B2	• Her zaman durumu uygun yapamasa bile, konuşmayı başlatabilir, uygun olduğunda söz alabilir ve gerektiğinde görüşmeyi sonlandırabilir
5/6	B1	• Basit ve doğrudan bir konuşmayı başlatıp, sürdürür ve bitirebilir. Karşılıklı anlamayı kesinleştirmek için karşısındakinin söylediklerinden bazı kısımları tekrarlayabilir.
3/4	A2	• Soruları yanıtlayabilir, aynı zamanda basit ifadelere tepki verebilir. Ne zaman anladığını belirtebilir, ancak konuşmayı kendisi sürdürebilecek kadar anlamaz.
1/2	A1	• Basit biçimde anlaşabilir, ancak iletişim tamamen konuşulanın yavaş tekrarlanmasına, farklı tanımlanmasına ya da düzeltilmesine bağlıdır.

0	• İletişim mümkün değildir • Notlandırılacak kadar dilsel üretim yapmaz