UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Linguistic competences of learners of Dutch as a second language at the B1 and B2 levels of speaking proficiency of the Common European Framework of Reference for Languages (CEFR)

Hulstijn, J.H.; Schoonen, R.; de Jong, N.H.; Steinel, M.P.; Florijn, A.

DOI
10.1177/0265532211419826

Publication date
2012

Document Version
Final published version

Published in
Language Testing

Link to publication

Citation for published version (APA):
Hulstijn, J. H., Schoonen, R., de Jong, N. H., Steinel, M. P., & Florijn, A. (2012). Linguistic competences of learners of Dutch as a second language at the B1 and B2 levels of speaking proficiency of the Common European Framework of Reference for Languages (CEFR). *Language Testing*, *29*(2), 203-221. https://doi.org/10.1177/0265532211419826

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:09 Mar 2023

Language Testing
http://ltj.sagepub.com/

Linguistic competences of learners of Dutch as a second language at the B1 and B2 levels of speaking proficiency of the Common European Framework of Reference for Languages (CEFR)

Jan H. Hulstijn, Rob Schoonen, Nivja H. de Jong, Margarita P. Steinel and Arjen Florijn

Language Testing 2012 29: 203 originally published online 28 November 2011
DOI: 10.1177/0265532211419826

The online version of this article can be found at:
http://ltj.sagepub.com/content/29/2/203

Published by:
SAGE
http://www.sagepublications.com

Additional services and information for *Language Testing* can be found at:

Email Alerts: http://ltj.sagepub.com/cgi/alerts

Subscriptions: http://ltj.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

Citations: http://ltj.sagepub.com/content/29/2/203.refs.html

>> Version of Record - Apr 23, 2012

OnlineFirst Version of Record - Nov 28, 2011

What is This?

Article

LANGUAGE
TESTING

Language Testing
29(2) 203–221
© The Author(s) 2011
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0265532211419826
ltj.sagepub.com
SAGE

Linguistic competences of learners of Dutch as a second language at the B1 and B2 levels of speaking proficiency of the Common European Framework of Reference for Languages (CEFR)[1]

Jan H. Hulstijn and Rob Schoonen
University of Amsterdam, the Netherlands

Nivja H. de Jong
Utrecht University, the Netherlands

Margarita P. Steinel and Arjen Florijn
University of Amsterdam, the Netherlands

Abstract

This study examines the associations between the speaking proficiency of 181 adult learners of Dutch as a second language and their linguistic competences. Performance in eight speaking tasks was rated on a scale of communicative adequacy. After extrapolation of these ratings to the Overall Oral Production scale of the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001), 80 and 30 participants (on average per speaking task) were found to be, respectively, at the B1 and B2 levels of this scale. The following linguistic competences were tapped with non-communicative tasks: productive vocabulary knowledge, productive knowledge of grammar, speed of lexical retrieval, speed of articulation, speed of sentence building, and pronunciation skills. Discriminant analyses showed that all linguistic competences, except speed of articulation, discriminated participants at the two levels of oral production. Subsequent comparisons showed that the distance between B1ers and B2ers was smaller in knowledge of high-frequency words than in knowledge of medium- and low-frequency words. Extrapolation from scores on the vocabulary test yielded estimations of productive vocabularies of, on average, 4000 and 7000 words for B1ers and B2ers, respectively. The grammar test assessed grammatical

---

Corresponding author:
Jan H. Hulstijn, Amsterdam Center for Language and Communication, University of Amsterdam, Spuistraat 134, 1012 VB Amsterdam, the Netherlands
Email: j.h.hulstijn@uva.nl

knowledge in 10 domains. B2ers were found to outperform B1ers on all parts of the test. Thus, the differences in lexical and grammatical knowledge of B1ers and B2ers appear to be a matter of degree, rather than a matter of category or domain. The paper ends with a research agenda for a linguistic underpinning of the CEFR.

Keywords

CEFR, grammar, levels of proficiency, linguistic competence, vocabulary

The Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001) contains proposals for formulating functional learning targets for language learning, teaching and assessment. Throughout Europe and beyond, the CEFR has become a major document of reference for language in education, with both the ambition and the potential of bringing common standards and transparency, across Europe, to the formulation of objectives of foreign-language learning curricula and the certification of foreign-language proficiency skills of citizens continuing their educational or professional careers in other European countries. The CEFR distinguishes between ‘language activities’ (p. 57) and ‘communicative language competences’ (p. 108). Communicative competences are subdivided into linguistic, sociolinguistic and pragmatic competences, apart from various kinds of nonverbal competences.

As Alderson (2007) and Hulstijn (2007, 2011) have pointed out, there is no evidence in terms of learner performance that a learner at a given level of the Overall Oral Production activity scale (p. 58) necessarily possesses linguistic competences at the same level (e.g. Vocabulary Range, Grammatical Accuracy, and Phonological Control, pp. 112, 114, and 117, respectively). In general, one might expect that performance at a given level along an activity matches linguistic competences at the same level, although the CEFR acknowledges the existence of so called uneven profiles (p. 17).[2] The CEFR levels, in their present form, are neither based on empirical evidence taken from L2-learner performance, nor on any theory in the fields of linguistics or verbal communication, as has been pointed out by, among others, Weir (2005), Alderson (2007), Alderson, Clapham, and Wall (2006), and Hulstijn (2007, 2011). Furthermore, as Hulstijn, Alderson, and Schoonen (2010, p. 17) argue, ‘what the CEFR does not indicate is whether learner performance at the six functional levels as defined in Chapter 4 actually matches the linguistic characteristics defined in Chapter 5, and, more specifically, *which linguistic features (for a given target language) are typical of each of the levels*’ (italics in original). The present study should be seen as a modest attempt to investigate this issue empirically. Its aim is to examine how well several linguistic competences discriminate learners who performed at the B1 and B2 levels of the Overall Oral production activity scale. We wondered if and how L2 learners at B1 and B2 levels of speaking proficiency differed in knowledge of vocabulary and grammar. Do learners at the B1 level of speaking proficiency (henceforth B1ers) simply know fewer words than learners at the B2 level of speaking proficiency (B2ers) across the whole lexicon, regardless of word frequency, or might it be that lexical knowledge of B1ers is restricted to high-frequency words, whereas B2ers know many low-frequency words in addition to high-frequency words? Similarly, in the domain of grammar, is grammatical knowledge of B1ers restricted to common phenomena, taught in beginner courses, and do B2ers know more phenomena (content difference), or is it the case that B1ers and B2ers only differ in the extent to which they are familiar with grammatical phenomena, regardless of how commonly they occur (degree difference)? Although there is abundant evidence in the literature that knowledge of vocabulary and grammar is strongly associated with L2 reading comprehension, writing ability, listening comprehension and speaking proficiency (e.g. Alderson, 2005; Bachman & Palmer, 1982; Fouly, Bachman, & Cziko, 1990; Harley, Cummins, Swain, & Allen, 1990; Milton, 2009, 2010; Shiotsu & Weir, 2007; Stæhr, 2008), to our knowledge these questions have not been previously investigated with respect to adjacent CEFR levels, although interest in this issue is rising (Bartning, Martin, & Vedder, 2010).[3]

We administered eight speaking tasks to 181 adult learners of Dutch as a second language (L2) to assess their speaking proficiency, that is, the adequacy with which participants were able to perform communicative speaking tasks at the A2, B1 and B2 levels of the Overall Oral Production activity scale, and a number of non-communicative tasks to separately assess their linguistic competences (knowledge and skills). Three types of linguistic competences were assessed: (1) declarative knowledge was assessed with tests of productive vocabulary knowledge and productive knowledge of grammar; (2) speed-of-processing skills were assessed with tests of speed of lexical retrieval, speed of articulation, and speed of sentence building; (3) pronunciation skills were assessed with a pronunciation test comprising measures of the quality of vowels, diphthongs, consonants, intonation, and word stress. In each of the eight speaking tasks, a fair number of learners was found to perform at the B1 or B2 level. We used discriminant analyses to examine how well the linguistic competences just mentioned (*not* transformed into CEFR levels) discriminated speaking proficiency at the B1 and B2 levels. We then examined knowledge of vocabulary and grammar, as measured by our vocabulary and grammar tests.

Our study addressed the following research questions:

1. How well do knowledge of vocabulary and grammar, speed of lexical retrieval, speed of articulation, speed of sentence building, and pronunciation skills discriminate learners of Dutch L2 at the B1 and B2 levels of speaking proficiency, as assessed in eight communicative tasks?
2. How do learners of Dutch L2 at the B1 and B2 levels of speaking proficiency differ in their vocabulary knowledge? Do they differ only in their knowledge of low-frequency words or do they differ in their knowledge of words in all frequency classes to the same degree?
3. How do learners of Dutch L2 at the B1 and B2 levels of speaking proficiency differ in their knowledge of grammar? Do B1ers lag behind B2ers in all morphosyntactic domains tested (degree difference) or is it the case that B1ers have mastered knowledge in some domains to the same degree as B2ers and lag behind B2ers only some other domains (content difference)?

Method

This section first provides information on the data collected from the L2 learners. A more comprehensive report is provided in De Jong, Steinel, Florijn, Schoonen, and Hulstijn (forthcoming, 2012). The second and third subsections report, respectively, how we established the link between the CEFR and the speaking tasks and between the CEFR and participants' performance in the speaking tasks (speaking proficiency).

Data collected from the L2 learners

Participants. Data were collected from 208 adult L2 learners of Dutch, of whom 181 were able to complete all tasks (age range 20–56 years [$M = 29$; $SD = 6$]; 72% female; 46 different first languages; length of residence in the Netherlands between 10 months and 20 years).

Assessment of speaking proficiency. (1) *Tasks and materials:* Speaking proficiency was measured with eight computer-administered speaking tasks, which required participants to look at the computer screen (providing the speaking cues) and to talk into a microphone. The eight tasks were constructed with contrasts on the following three dimensions, in a 2 × 2 × 2 fashion: complexity of the topic (complex versus simple), formality of the setting (informal versus formal) and discourse type (descriptive versus argumentative). The task instructions specifically mentioned the audience that participants should address in each task and requested participants to 'role play' as if they were actually speaking to these audiences. For each task, the instruction screens provided a photo picture of the communicative situation and one or several visual-verbal cues concerning the topic. The tasks are described as follows:

> *Apartment* (simple, informal, descriptive): Participant speaks on the phone to a friend, describing a common friend's new apartment.
> *Road accident* (simple, formal, descriptive): Participant, who has witnessed a road accident some time ago, is in a courtroom, describing to the judge what had happened.
> *Advice* (simple, informal, argumentative): Participant advises his or her sister on how to choose between (or combine) child care, further education, and paid work.
> *Playground* (simple, formal, argumentative): Participant is present at a neighborhood meeting in which an official has just proposed to build a school playground, separated by a road from the school building. Participant raises her or his hand and argues against the planned location of the playground.
> *Unemployment* (complex, informal, descriptive): Cued by a graph, participant tells a friend about the development of unemployment among women and men over the last ten years.
> *Hospital* (complex, formal, descriptive): Participant works at the employment office of a hospital and tells a candidate for a nurse position what the main tasks in the vacant position are.
> *Transportation* (complex, informal, argumentative): Participant discusses the pros and cons of three means of transportation (public transportation, bicycle, and automobile) on how to solve the problem of traffic congestion.

> *Car park* (complex, formal, argumentative): Participant is manager of a supermarket, addressing a neighborhood meeting, arguing which one of three alternative plans for building a car park he or she prefers.

(2) *Rating of speaking proficiency:* Participants' speaking proficiency was measured in terms of the communicative adequacy (CA) of their responses in the eight speaking tasks, as rated by a panel of 12 judges, such that each response was rated by four judges. The scales were built on the same pattern, comprising six levels, containing (task-specific) descriptors pertaining to (a) the amount and detail of information conveyed, relevant to the topic, setting (formal/informal) and discourse type (descriptive/argumentative) and (b) the intelligibility of the response. Note that the scale did *not* contain references to linguistic quality. To allow for an even more precise distinction between responses, each of the six levels was subdivided into five sublevels, resulting in a rating scale ranging from 1 to 30. Responses rated 1 to 5, 6 to 10, and 11 to 15 were described as being insufficient in terms of communicative adequacy, with descriptors such as 'unsuccessful', 'weak', and 'mediocre'. Responses rated 16 to 20, 21 to 25 and 26 to 30 were described as being sufficient in terms of communicative adequacy, with descriptors such as 'sufficient', 'quite successful', and 'very successful'.

After an introductory training session, the judges received all responses of either two or three tasks, such that for each speaking task, four judges independently rated all speaking performances, randomized per task and rater.

The linguistic-competence tasks. (1) *Vocabulary knowledge:* For the assessment of productive vocabulary knowledge, a paper-and-pencil task was administered, consisting of two parts.[4] Part 1 (90 items) elicited knowledge of single words; part 2 (26 items) elicited knowledge of multi-word units (collocations). For part 1, nine words were selected from each frequency-band of 1000 words between words ranked 1 to 10.000 according to the Corpus of Spoken Dutch (CGN) (Nederlandse Taalunie, 2004). We used the format suggested by Laufer and Nation (1999): for each item a meaningful sentence was presented with the target word omitted, except for its first letter(s). Part 2 of the vocabulary task tested knowledge of 26 prepositional phrases and verb-noun collocations; the preposition or main verb was omitted and the gap had to be filled in. The test format was the same as in the first part of the vocabulary test, except that no first letter(s) was given. For each correct response, one point was awarded.

(2) *Grammar knowledge:* The grammar task consisted of 142 items covering a range of grammatical issues grouped in different test sections. At the beginning of each section, short instructions and an example were given. Knowledge of the following types of grammatical features was assessed: Inflectional variants of verbs (19 items) and adjectives (19 items), word order in main clauses and subclauses (25 items), the place of particles of so called particle verbs (eight items), dummy pronouns (26 items), order of modal adverbs in sentences with more than one adverb (10 items), relative pronouns (15 items), possessive pronouns (five items), choice of auxiliary verbs (10 items), and construction of passive sentences (five items). For each correct response, one point was awarded.

(3) *Lexical retrieval speed:* This was tested with a computer-administered picture-naming task. Participants saw 28 pictures one-by-one and named the corresponding word as quickly as they could. The words corresponding to the pictures belonged to the 2200 most frequent lemmas in the CGN. A script written in PRAAT (Boersma & Weenink, 2005) automatically measured the time between the appearance of the pictures and the beginning of the correct responses in milliseconds (ms).

(4) *Speed of articulation: response latency and response duration:* In order to elicit well-prepared articulation, and measure its speed, participants carried out the picture naming task once more. This time, however, they were asked to prepare their response to naming a picture but wait with the actual naming of the picture until a cue was given. Response latency was measured as the latency between the auditory cue and the beginning of the response. Response duration was measured as the duration of the response, that is, the latency between the beginning and the end of the response.

(5) *Sentence building speed:* This was assessed with a computer-administered sentence completion task, in which the completion of the sentence was an alteration of a given sentence. For instance, participants would hear and read ‘De meisjes gaan meestal naar de bakker.’ (*The girls usually go to the bakery*), after which the written cue ‘Het meisje ...’ (*The girl ...*) was presented. The correct response would be ‘Het meisje gaat meestal naar de bakker’ (*The girl usually goes to the bakery*). The alteration of the sentences always involved a grammatical change that was induced by the written cue following the original sentence. Grammatical changes required adjectival inflection (10 items), verbal inflection of number (10 items), verbal conjugation changing present tense to past tense (10 items), construction of subclauses from main clauses (10 items), or subject-verb inversion in main clauses (10 items). The period between the beginning and the end of the participant’s response was measured (in ms).

(6) *Pronunciation:* This was assessed with a computer-administered task. Sixty mostly monosyllabic target words were selected, covering a broad range of vowels, diphthongs and consonants, each sound occurring in one to four target words. Thirty-six of these words were divided into six sets of six single-word items, and the 24 remaining target words were embedded in 15 sentences. Ten of these sentences were also designated to test the quality of the intonation pattern. Finally, to test word-stress knowledge, 10 words with two to four syllables were added, divided into two sets of five target words. Each response was rated by three students in phonetic sciences (see De Jong et al., forthcoming 2012, for details).

Classification of the speaking tasks

In the present study, the transformation of the communicative-adequacy ratings of responses in the eight speaking tasks to the classification of our participants at the B1 or B2 speaking proficiency levels of the CEFR, consisted of several steps. To the extent applicable, we used the guidelines given in the pilot version of the Manual *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)* (Council of Europe, 2003). Relating a language test or exam to the CEFR is a complicated matter, as has been pointed out by several experts (e.g. Alderson, 2005; Alderson, Figueras, Kuijpers, Nold, Takala, & Tardieu, 2006; Figueras, North, Takala, Verhelst, & Van Avermaet, 2005; Tannenbaum & Wylie, 2005). We could not follow all procedures contained by the Manual because our eight speaking tasks do not constitute an exam nor does our study pertain to the validation of a system of language exams or certificates.

The expert raters. We invited six experts who had been previously involved in CEFR standard setting studies. Of these six experts, three were test developers by profession, working with a Dutch-Flemish certification organization; their experience in test development ranged between five and 20 years. Of the other three experts, one had been involved for more than two years in the production of the state examinations Dutch as a second language, and two experts had experience of 13 and 25 years in teaching L2 Dutch, developing the Dutch L2 exam used by Dutch universities (the exam system which preceded the state examinations), and rating test takers' performance.

Procedure. The rating session lasted 3.5 hours, intermissions not included. With respect to the purpose of the meeting, participants were told that we needed their expert judgment on the CEFR levels of the eight speaking tasks of our study. Furthermore, we wanted their judgment on a small sample of responses. Accordingly, the session consisted of two parts: task linking and response linking. The latter part served as a warm up for three of the experts, who had agreed to rate 360 speaking responses, after the session. We report on the second part in the subsection 'Rating of speaking proficiency on the CEFR scales'.

After the experts had filled out a questionnaire concerning their experience in language teaching and testing, they were asked to refresh their familiarity with the CEFR by reading parts of chapters 4 and 5 of the CEFR, relevant to speaking assessment. Although we used the official Dutch translation of the CEFR (Nederlandse Taalunie, 2006), we refer here to the original English scales, as published by the Council of Europe (2001). The experts were then shown the eight speaking tasks. Next, they performed two activities designed to make them familiar with the nature of the eight speaking tasks before assigning each task a CEFR level. In the first task-familiarization activity, they wrote down which domains (personal, public, occupational, and educational; Council of Europe, 2001, pp. 48–49) were applicable for each of the eight speaking tasks. In a subsequent discussion, participants agreed that it is the setting rather than the topic that determines the domain. They then agreed that the Apartment, Advise, Unemployment, and Transportation tasks had to be placed in the personal domain; the Road Accident, Playground and Car Park tasks were placed in the public domain, while the Hospital task was classified in the professional domain.

In the second task-familiarization activity, the participants wrote down how they would describe the 'communicative task' (Council of Europe, 2001, pp. 53–55) of each speaking task. This resulted in descriptions varying in detail, but all minimally containing a do verb with a noun phrase or an adverbial phrase, for example 'to give one's opinion concerning the location of a playground' and 'to provide arguments which location is the best one for a playground, during a neighborhood meeting, as a member of the audience'.

In the following step, the experts placed the eight speaking tasks on the following three scales: (1) *Sustained Monologue: Describing Experience*, (2) *Sustained Monologue: Putting a Case*, and (3) *Addressing Audiences* (Council of Europe, 2001, pp. 59–60).

They were allowed to write 'not applicable' if deemed appropriate. All experts then told how they had classified the tasks on these three scales and a discussion followed. Participants agreed that four tasks could be placed on the scale *Sustained Monologue: Describing Experience* (Apartment, Road Accident, Unemployment, and Hospital), that four tasks could be placed on the scale *Sustained Monologue: Putting a Case* (Advise, Playground, Transportation, Car Park), and that one task could also be placed on *Addressing Audiences* (Car Park). A large consensus on the CEFR level of the tasks emerged already at this stage. In the final activity in the first part of the session, the experts gave their rating of the eight speaking tasks on the scale *Overall Oral Production*, adding the motivation for their ratings.

Table 1 shows the task features, as originally planned and designed in terms of formality, discourse type, and topic complexity, compared to the expert ratings on the CEFR scale *Overall Oral Production*. The three design features of the Apartment task and those of the Car Park task, taken together, make these two tasks the least and most complex task respectively, which is clearly reflected in the CEFR ratings, that is, unanimously A2 and B2, respectively.

Rating of speaking proficiency on the CEFR scales

Training of the expert raters. In the second part of the session, the experts listened to recordings of eight speaking responses, one taken from each task. Four responses had received relatively low, and four responses had received relatively high communicative-adequacy ratings (see subsection 'Rating of speaking proficiency'). Each response was played twice. During and after listening, the experts rated the responses on the following nine scales: *Overall Oral Production, Vocabulary Range, Vocabulary Control, Grammatical Accuracy, Phonological Control, Sociolinguistic Appropriateness, Cohesion and Coherence, Spoken Fluency, and Propositional Precision* (Council of Europe, 2001, pp. 58, 112, 114, 117, 122, 125, and 129). The purpose of this part of the session was to train

Table 1. The eight speaking tasks by task type and expert rating

Task	Formality	Discourse type	Topic Complexity	Majority task rating on CEFR scale *Overall Oral Production*	Number of experts awarding the majority rating (Max. = 6)
Apartment	Low	Descriptive	Low	A2	6
Road Accident	High	Descriptive	Low	B1	6
Advice	Low	Argumentative	Low	B2	4 (5)*
Playground	High	Argumentative	Low	B1	4
Unemployment	Low	Descriptive	High	B1	5
Hospital	High	Descriptive	High	B2	6
Transportation	Low	Argumentative	High	B2	5 (6)*
Car Park	High	Argumentative	High	B2	6

*One rater was undecided between B2 and B1.

the three experts who had agreed to rate a large subsample of all responses (see below), in interpreting and using the nine scales. Not surprisingly, the experts found it initially difficult to rate responses on so many scales simultaneously. After listening to and rating each response, the experts engaged in lively discussions on the subtle differences between the scale values and on the possibilities of assigning different ratings on different scales to a response, when the response reflected what the Manual calls a learner's 'unequal profile' (Council of Europe, 2009, p. 43). From the third response onwards, participants were able to give their ratings on all nine scales, during listening (twice).

Three of the experts had agreed to rate a large sample of responses (as specified in the next subsection) on the aforementioned nine scales. Having signed a declaration of confidentiality, they received an audio CD with all responses, electronic/paper rating forms, and a list with instructions. They rated the responses at their homes and returned all materials within a few weeks.

Selection of responses. A subsample of the responses that had been rated for communicative adequacy (see subsection 'Rating of speaking proficiency') were re-rated on the CEFR scales by the expert raters. The responses to be re-rated were selected in the following way. For the (original) CA ratings, we computed the mean (*M*) and standard deviation (*SD*) of the four ratings per response. Thus, the mean indicated the level of the response and the standard deviation the agreement among the four raters. For each task, we then selected between 43 and 45 responses in such a way that the proportional distribution across the score scale reflected the distribution of the whole set of responses. We had divided the scale into six parts of the distributions: between 1 and 2 *SD* from the mean level (at both sides of the mean), between 0.5 and 1 *SD* from the mean (at both sides), and between the mean and 0.5 *SD* from the mean (at both sides). Furthermore, we selected clear cases, that is, responses that were rated with a high level agreement (i.e. a small *SD* for the four ratings). Note that the responses were selected for each task independently.

Extrapolation. Three experts evaluated the selected responses in the eight speaking tasks on the CEFR's main activity scale *Overall Oral Production*. In 84% of the responses (range across tasks: 71–96%), all three experts gave the same rating or two experts gave the same rating and one expert gave a rating at an adjacent level. Cronbach's alpha was an index of internal consistency of the panel score, ranging between .74 and .84 across tasks. Cronbach's alpha probably slightly underestimates the reliability, because raters differed in the variance of their scores. In the remainder of this paper, we refer to this rating on the activity scale *Overall Oral Production* with the label 'CEFR ratings'.

As we had expected, the mean CEFR ratings were strongly associated to the CA ratings of the same 43–45 responses in all eight tasks (*r* between .81 and .88). To extrapolate the mean CEFR ratings of the re-rated responses to the CA scores of the entire sets of 181 responses, for each speaking task separately, we applied the regression method (cf. Engelen & Eggen, 1993). In the computations, decimal CEFR scores were used; the ultimate decimal scores were rounded.[5] On the basis of this extrapolation procedure, all responses could be classified as being at one of the six CEFR levels.

Results

This section is divided into subsections addressing the three research questions in turn.

Prediction of B1 and B2 speaking ratings

In this subsection we report the results of discriminant analyses, in which participants' scores in the linguistic-competence tasks were used to discriminate between responses in the eight speaking tasks, rated at the B1 or the B2 CEFR levels.

As mentioned above, we had targeted the study on L2 learners at intermediate levels of proficiency and had designed speaking tasks at appropriate levels of difficulty. We had therefore expected that, after the CA scores had been converted into CEFR scores, the vast majority of our 181 L2 learners would be found to be at the B1 and B2 levels. It turned out that hardly any response was at the A1 level (ranging from 0–1% per task), a small number was at the A2 level (6–18%), the majority was either at the B1 level (49–65%) or the B2 level (20–36%), a smaller number was at the C1-level (1–7%), and just one response was considered C2. These findings thus met our expectations. Because of the uneven distribution of responses across CEFR levels, we conducted discriminant analyses only for the two major categories, namely responses at the B1 and B2 levels, using subjects' scores on the linguistic-competence tests (called predictors) to explore the extent to which they discriminate between CEFR levels.

Prior to the discriminant analyses, we explored the B1–B2 group differences with univariate ANOVAs. These analyses showed that B1 and B2 performers differ in all linguistic competences, except articulation speed (response duration and response latency). In terms of effect sizes, the effects are large for most of the comparisons ($\eta^2 > .14$, cf. Cohen, 1988). Vocabulary knowledge showed the largest differences across all tasks.

Discriminant analyses were conducted for each speaking task separately, with just the participants at either one of the two levels. Across tasks, the numbers of subjects at B1 and B2 levels of speaking (called B1ers and B2ers) ranged between 89 and 111 (B1) and between 34 and 63 (B2). Each discriminant analysis leads to a discriminant function, that is, a weighed score of the predictors. This function can be interpreted by looking at the correlation between the weighed score and the predictors individually. Table 2 shows a summary of these correlations as found for each speaking task in terms of the median of the eight correlations and the min. and max. correlations. Note that negative values are expected with respect to the speed predictors because lower reaction times are associated with higher communicative-adequacy scores. Most correlations are substantial, except for the two articulation measures, as could be expected on the basis of the initial ANOVAs.

Another way of evaluating the results of the discriminant analyses is akin to conducting regression analysis. The question then is as follows: Are all predictors necessary to discriminate between B1ers and B2ers or can this be done in a more parsimonious way? When there is a high degree of collinearity between predictor variables, distinguishing between B1ers and B2ers will require fewer variables. Table 3 shows the results of a stepwise procedure to investigate which predictors are needed for an optimal and parsimonious prediction of B1ers or B2ers (the statistic criterion for entering a variable in the discriminant function was $F > 3.84$; the criterion for dropping a variable was $F < 2.71$).

Table 2. Median correlation and range, between discriminant function (across all eight speaking tasks) and predictor variables

Predictor variables	Median	Min.	Max.
Vocabulary	.860	.792	.890
Grammatical knowledge	.655	.620	.813
Pronunciation	.733	.628	.757
Sentence building	−.679	−.529	−.804
Lexical retrieval	−.486	−.404	−.555
Articulation latency	−.227	−.097	−.340
Articulation duration	−.006	−.075	.195

Table 3. Predictor variables remaining in stepwise analysis

Speaking tasks	Predictor variables
Advice	Vocabulary, Sentence building
Transportation	Vocabulary, Sentence building
Unemployment	Vocabulary, Pronunciation, Articulation latency
Accident	Vocabulary, Sentence building
Playground	Vocabulary, Lexical retrieval, Pronunciation
Hospital	Vocabulary, Sentence building
Car park	Vocabulary, Pronunciation, Lexical retrieval
Apartment	(Vocabulary),* Pronunciation, Lexical retrieval, Grammar

*In this analysis, Vocabulary knowledge is entered first, but removed in the final step when grammatical knowledge is entered. This is due to the multicollinearity between the variables.

Vocabulary knowledge is always the first and best predictor, but in all cases prediction improves when another variable is added, most often Sentence building and Pronunciation.

Of course it is also important to know whether there are any prediction differences between the speaking tasks. Do the linguistic predictors predict level of speaking equally well across speaking tasks, or are predictions better for complex tasks, formal tasks or, for example, B2-level tasks? Table 4 summarizes the percentages of successful classifications of participants as being B1 or B2, in relation to task features. The success rates range from 70% to 85%, while there is no evident systematic relationship between success rate and task features.

In summary, the answer to RQ 1 is that linguistic knowledge (knowledge of vocabulary and grammar), the speed with which linguistic knowledge can be processed (speed of lexical retrieval and speed of sentence building), as well as pronunciation skills were found to form important predictors of whether the adult L2 learners in this study were found to perform speaking tasks at the B1 or the B2 level of the CEFR (Tables 2 and 3), with vocabulary knowledge being the strongest predictor. Speed of articulation (in the articulation task) was not associated with ratings of overall speaking proficiency (in the speaking tasks). The classification of the eight speaking tasks in terms of formality, discourse type or topic complexity or in terms of rated CEFR level (A2, B1, B2) did not

Table 4. Successful classification of participants as B1 or B2 (in percentage, from low to high), by task and task features*

Task	Task CEFR rating	Formality	Discourse type	Topic complexity	%
Advice	B2	Low	Persuasive	Low	70
Playground	B1	High	Persuasive	Low	75
Car park	B2	High	Persuasive	High	76
Accident	B1	High	Descriptive	Low	77
Hospital	B2	High	Descriptive	High	79
Apartment	A2	Low	Descriptive	Low	79
Transport	B2	Low	Persuasive	High	80
Unemployment	B1	Low	Descriptive	High	85

*Classification success results from the comparison of the division of participants into B1 and B2 (on the basis of rated speaking performances) with the division as predicted by the Discriminant function (with linguistic predictors).

appear to play an important role in how well performance on the linguistic-competence tasks discriminated speaking proficiency rated at the B1 or B2 CEFR level (Table 4). On the basis of participants' performance in these predictor tasks one would be able to classify 70 to 85% of participants correctly at either level.

Comparisons of vocabulary and grammar knowledge of B1ers and B2ers

With research questions 2 and 3 we examine the differences in linguistic knowledge in the domains of lexis and grammar between L2 learners at the B1 and B2 levels of speaking proficiency. In this subsection we examine how vocabulary and grammar knowledge of B1ers and B2ers differed. To classify subjects as B1er or B2er in the analyses reported in this subsection we looked at their (extrapolated) CEFR scores in the eight speaking tasks. With at least six out of eight CEFR scores at the same level as the classification criterion, 80 B1ers and 30 B2ers emerged.

Vocabulary knowledge. The vocabulary test as a whole (116 items) was highly reliable for all 181 participants (alpha = .98), for the 80 B1ers (.95), and for the 30 B2ers (.95). B2ers performed substantially and significantly better than B1ers on all sections of the vocabulary test (Table 5). Research question 2 asked: Do B1ers simply know fewer words than B2ers across the whole lexicon, or might it be that lexical knowledge of B1ers is restricted to high-frequency words, whereas B2ers know many low-frequency words in addition to high-frequency words? The distance between B1ers and B2ers was somewhat bigger in the medium- and low-frequency words than in the high-frequency words (see Table 5). A repeated-measures ANOVA with Frequency (High, Medium, Low) as the within-subjects variable and Group (B1 vs. B2) as the between-subjects variable, yielded a significant main effect of Frequency ($F[2,216] = 448.367$; $p = .000$; $\eta^2_p = .806$), a significant main effect of Group ($F[1,108] = 77.996$; $p = .000$; $\eta^2_p = .419$), and a significant Frequency × Group interaction ($F[2,216] = 7.476$; $p = .001$; $\eta^2_p = .065$). The

Table 5. Performance of B1ers and B2ers on the productive vocabulary test, by test part*

Test part	CEFR	*N*	*k*	*alpha*	*M*	*SD*	95% CI	
							LL	*UL*
High	B1	80	30	.86	18.9	5.8	17.6	20.2
	B2	30	30	.83	26.1	3.9	24.6	27.5
Medium	B1	80	30	.87	11.5	6.0	10.2	12.8
	B2	30	30	.79	21.4	4.5	19.7	23.1
Low	B1	80	30	.80	5.4	4.1	4.5	6.3
	B2	30	30	.84	15.3	5.8	13.1	17.4
Total	B1	80	90	.94	35.8	14.6	32.6	39.0
	B2	30	90	.93	62.7	13.2	57.8	67.6
Collocations	B1	80	26	.84	8.8	4.8	7.7	9.9
	B2	30	26	.85	18.0	5.0	16.1	19.8

* Test parts: Part 1 single content words of high, medium, and low frequency (30 items each); Part 2 collocations (26 items).

effect sizes of the main effects are large, but the effect size of the interaction is medium. Subsequent univariate ANOVAs for each of the four test parts showed significant differences between B1ers and B2ers ($p < .000$) with η^2_p values of .264 (High), .384 (Medium), .479 (Low), and .418 (Collocations).

Thus, the answer to RQ2 is that it is not the case that B1ers' vocabulary knowledge is mainly restricted to high-frequency words. The B1ers knew some low-frequency words as well but the distance between their vocabulary knowledge and that of B2ers increases at lower word-frequency levels. An extrapolation of the mean correct responses of B1ers and B2ers on the first part of the test (90 single-word items) to knowledge of the 10,000 most frequent words in the corpus of spoken Dutch yields productive vocabularies of almost 4000 words (35.8/90 × 10,000 = 3977) for B1ers (with a 95% confidence interval between 3622 and 4333 words) and almost 7000 words (62.7/90 × 10,000 = 6966) for B2ers (with a 95% confidence interval between 6422 and 7511 words).

Grammar knowledge. The grammar test as a whole (142 items) was reliable for all 181 participants (alpha = .95; mean correct score = 107; *SD* = 20), for the 80 B1ers (alpha = .90; *M* = 101; *SD* = 15), and for the 30 B2ers (alpha = .77; no errors on 27 items; *M* = 122; *SD* = 8).

Not surprisingly, the B2ers performed significantly better than the B1ers on the test as a whole (86% over 71%) as well as on all parts of the test (Table 6). A 10 × 2 repeated-measures ANOVA with Linguistic domain as the within-subjects variable and Group (B1 vs. B2) as the between-subjects variable, yielded a significant main effect of Linguistic domain (F[9,972] = 21.007; $p = .000$; $\eta^2_p = .163$), a significant main effect of Group (F[1,108] = 50.493; $p = .000$; $\eta^2_p = .319$), and a significant Linguistic domain × Group interaction (F[9,972] = 4.015; $p = .000$; $\eta^2_p = .036$). The effect sizes of the main effects are large, but the effect size of the interaction is between small and medium.

As can be gleaned from Table 6, B1ers performed more poorly than B2ers on all sections of the test (hence the main group effect), but the effect sizes differ substantially by linguistic domain (hence the significant Linguistic domain × Group interaction).

Table 6. Performance (% correct) of B1ers and B2ers, by linguistic domain*

Linguistic domain	*k*	B1ers (*n* = 80)		B2ers (*n* = 30)		Diff.	η^2_p
		Mean	SD	Mean	SD		
Verb forms	19	80	17	94	7	14	.149
Adjective inflection	19	76	14	84	12	8	.068
Word order	25	81	11	91	8	10	.154
Particle verbs	8	66	29	83	14	17	.087
Dummy pronouns	26	69	10	81	7	12	.263
Modal-adverb order	10	57	19	84	16	27	.319
Relative pronouns	15	56	22	83	12	27	.275
Possessive pronouns	5	86	17	98	6	12	.127
Auxiliary choice	10	66	20	84	14	16	.158
Passive	5	59	32	81	29	22	.090

*k = number of test items; Diff. = difference between the two means; η^2_p = effect size of univariate tests, comparing the two means for each subtest separately (p = .006 or smaller, in all comparisons).

Effect sizes were relatively large in most domains, in particular in the domains of order of modal adverbs, relative pronouns, and dummy pronouns, topics usually not covered in teaching materials for beginners. In contrast, B1ers performed relatively well on verb forms, word order in main clauses and subclauses, and possessive pronouns, topics covered in all textbooks for beginners (e.g. Boers et al. 2004). But even in these domains B1ers lagged behind B2ers. Hence there is no evidence suggesting that the B1ers had mastered grammatical form-function mappings in any of the grammatical domains assessed. In contrast, the B2ers performed well in all sections of the test (between 81% and 98% correct), suggesting that many of them had mastered some or most of the grammatical phenomena assessed. The data do not provide hard evidence, however, because the test sections could not cover all phenomena in their domains and were not necessarily equally difficult. In sum, the evidence concerning RQ3 suggests that the difference in knowledge of grammar between the average B1er and the average B2er is more a matter of degree than of content.

Discussion

In the Introduction, we raised the question to what extent differences in speaking proficiency, assessed in eight speaking tasks, at two adjacent levels on the Overall Oral Production scale of the CEFR (B1 and B2) are associated with differences in linguistic competence. The study reported here aimed at looking into this question by examining speaking proficiency and linguistic competences of 181 adult learners of Dutch as a second language. All linguistic-competence predictors except articulation speed (latency and duration) were observed to discriminate the B1 and B2 levels of speaking proficiency (Tables 2 and 3). In terms of effect sizes, the effects are large for most of the comparisons ($\eta^2 > .14$, cf. Cohen, 1988), the strongest predictor being vocabulary knowledge. Although the eight speaking tasks differed in formality, discourse type and topic complexity, these differences did not affect the success with which participants could be correctly classified as being at the B1 or B2 level on the basis of their performance in the linguistic-competence tasks (Table 4).

Subsequently, the lexical and grammatical knowledge of 80 B1ers and 30 B2ers were analyzed in some detail. With respect to performance on the paper-and-pencil productive vocabulary test, we observed, as expected, that the distance between participants at the B1 and B2 levels of speaking proficiency was smaller in knowledge of the high-frequency words than in knowledge of the medium- and low-frequency words. Extrapolation from vocabulary scores yielded estimations of productive vocabularies of 4000 and 7000 words for B1ers and B2ers, respectively. The paper-and-pencil productive grammar test (142 items) assessed grammatical knowledge in 10 domains. B2ers were found to outperform B1ers on all parts of the test (on average, 86% over 71% correct), but the distances were more pronounced in the domains not covered in teaching materials for beginners (order of modal adverbs, dummy pronouns, and relative pronouns) than in domains taught to beginners (verb forms, word order in main clauses and subclauses, and possessive pronouns).

Given the documentation in the field of Dutch L2 curriculum development and testing (Bossers, 2010), and given the lexical and grammatical targets taught in widely used Dutch L2 textbooks (e.g. Boers, 2004), we had expected B1ers to have productive control in most of the grammar domains tested (especially verb forms and word order), and we had expected B1ers to productively control around 3000 words. The findings of this study generally provide support for these expectations. The 4000 figure for vocabulary at the B1-level (albeit with a 5% lower bound of 1667 words, that is, ignoring the 5% lowest scores) is much higher than the approximately 2000 words recommended for French (Coste, Courtillon, Ferenczi, Martins-Baltar, & Papo, 1987) and English (Van Ek & Trim, 1991) at the Threshold level, the predecessor of the CEFR B1 level, although the 2000 figure should perhaps best be understood as a minimal target. Nation (2001, p. 17) found that the 2000 most frequent words of English covered 90% of the words in a corpus of oral conversation (lower percentages of coverage were found in fiction, newspapers and academic text: 86%, 80%, and 78% respectively). To our knowledge, no such coverage figures exist for Dutch. The participants in the present study, who were preparing for enrollment at a university, were obviously aiming at a vocabulary target well beyond what is required for everyday oral conversation. This might explain why their vocabulary knowledge was ahead as it were of their speaking proficiency (what the CEFR calls an uneven profile); but this is just speculation. Further research is needed to establish whether knowledge of around 2000 words (depending on the way in which the language under study combines morphemes to form words and renders the resulting words between spaces in writing) does indeed suffice for speaking at the B1 level. Furthermore, our finding that the average B1er has knowledge of the inflection of most frequent verbs, and of the place of the finite verb in main clauses and subclauses (sections 1 and 3 of the grammar test; see Table 6) needs to be replicated in studies examining learners of other, typologically similar languages.

A limitation of our study is that we assessed speaking proficiency only with computer-administered tasks, albeit clearly framed in interactive communicative settings, contrasting in topic complexity, setting formality, and discourse type. Thus, strictly speaking, ‘spoken interaction’, categorized as a separate type of oral communication in the CEFR (Council of Europe, 2001, pp. 73–82), was included in this study only in a limited sense.

With our study we attempted to contribute to a critical examination of some implications of the CEFR, relevant to the field of language testing. Clearly, more work is needed. The final section of this paper is devoted to what we see as a research agenda.

A research agenda for underpinning the CEFR

As our study represents only a modest attempt to establish the linguistic competences of L2 users at two CEFR levels in one particular language (see the Introduction), we propose the following research agenda. First, as one of the CEFR authors stated (B. North, personal communication with the first author, April 16, 2010), what the CEFR calls ‘uneven profiles’ (Council of Europe, 2001, p. 17; Council of Europe, 2009, p. 43) are the rule and what the manual calls ‘flat profiles’ (Council of Europe, 2009, p. 43) are the exception. If this is so, research is needed into the extent to which profiles can actually be uneven. How linguistically imperfect (in terms of vocabulary, grammar, pronunciation/intonation, articulation speed) can performance on a C1 task be without failing as a communicative act, and to what extent can weaknesses in one component of linguistic competence be compensated with strengths in another component at a given CEFR level? These questions appear to be particularly relevant for the higher levels (B2, C1, and C2).

Second, research is needed on how little linguistic competence is minimally required to perform tasks at the lower levels (A1, A2, and B1). Vocabulary appears to be the most important linguistic component at the lower levels. Milton (2009, pp. 185–192) presents data from Hungarian learners of English as a foreign language at the A2 and B1 levels, with average vocabulary sizes of 2156 and 3264, respectively. English, Greek and Spanish learners of French as a foreign language were found to have average vocabulary sizes of 850, 1640 and 1700, respectively, at the A2 level, and of 850, 2422 and 2194 at the B1 level (see also Milton & Alexiou, 2009). Widely divergent figures are also given in Decoo (2011, chapter 5). The B1 figures are all substantially lower than the average 4000 words at the B1 level that we obtained. Such disparities point to the need of more empirical research and agreement on methods of establishing language proficiency at a given CEFR level, methods of establishing word knowledge, and ways of making proper cross-language comparisons. In terms of grammatical knowledge, the question remains which grammatical and phonotactic elements a learner must minimally control at given CEFR levels in the case of typologically divergent languages. Note that research on these questions is particularly needed in the productive skills (speaking and writing).

The question of what learners need to know in the domains of listening and reading appear to be somewhat less pertinent because receptive knowledge minimally required can be deduced from the analyses of corpora of language spoken and written in the activities specified in Chapter 4 of the CEFR. However, there as well, it would be relevant to conduct research using real L2 learners to find out which linguistic skills are minimally needed to understand speech and text in these activities (obviously, L2 learners ideally understand everything).

Findings from the empirical research proposed here might support or be at variance with the findings of the 'profiling' work, ongoing under the auspices of the Council of Europe (URL: http://www.coe.int/t/dg4/linguistic/dnr_EN.asp). This work consists of specifying, on the basis of the experience of curriculum designers, teachers and language testers, the precise content of each CEFR level for every language (so called 'reference level descriptions'). Thus, empirical research on the linguistic underpinnings of the CEFR's dimensions of activities should be placed high on the agenda of the Council of Europe and the language testing community.

Acknowledgements

This research was funded by the Netherlands Organisation for Scientific Research by a grant awarded to Hulstijn and Schoonen (NWO grant 254-70-030). We thank our research assistants Renske Berns, Andrea Friedrich, and Kimberley Mulder. We thank Ton Wempe and Rob van Son for their technical support and advice. We thank the anonymous reviewers for their useful comments.

Notes

1. The study reported here forms part of the What Is Speaking Proficiency (WiSP) project, conducted at the University of Amsterdam. Some of the findings, not related to the CEFR, are reported in De Jong, Steinel, Florijn, Schoonen and Hulstijn (forthcoming, 2012), concerning the componential nature of L2 learners' L2 skills, and in De Jong, Steinel, Florijn, Schoonen and Hulstijn (forthcoming, 2011), concerning the effect of task complexity on native and non-native speakers' functional adequacy, fluency, and lexical diversity.
2. There is no room here to do justice to the debates in many countries with respect to the virtues and vices of the CEFR. See, for instance, the contributions to *The Modern Language Journal*, 91(4) (2007) pp. 641–685; Alderson, Figueras, Kuijpers, Nold, Takala, and Tardieu (2006); Davidson and Fulcher (2007); and Jones and Saville (2009).
3. Dávid (2007) examined how performance on different item types in a grammar test differed between Hungarian learners of L2 English at different CEFR levels of proficiency. However, the study did not aim at identifying the contents of grammatical knowledge at different CEFR levels.
4. We concur with Read (2000, p. 157) that being able to provide the right word in what might be called a recall task is not the same as being able to use that word correctly and appropriately in one's speech or writing.
5. One could argue that truncation is more appropriate, because, for example, B2 level (score 4 on the six point scale) is only reached after the language learner has passed a certain threshold (score 4). However, we consider our scores as approximations of a language learner's level and then rounding to the nearest integer seems fairer.

References

Alderson, J. C. (2005). *Diagnosing foreign language proficiency*. London: Continuum.

Alderson, J. C. (2007). The CEFR and the need for more research. *The Modern Language Journal*, *91*, 659–663.

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

Alderson, J. C., Figueras, N., Kuijpers, H., Nold, G., Takala, S., & Tardieu, C. (2006). Analysing Tests of Reading and Listening in relation to the CEFR: The experience of the Dutch CEFR Construct Project. *Language Assessment Quarterly*, *3*, 3–30.

Bachman, L. F., & Palmer, A. S. (1982). Convergent and discriminant validation of oral language proficiency tests. *TESOL Quarterly*, *16*, 449–465.

Bartning, I., Martin, M., & Vedder, I. (Eds.). (2010). Communicative proficiency and linguistic development: Intersections between SLA and language testing research. *Eurosla Monograph Series*, *1*. Retrieved from http://eurosla.org/monographs/EM01/EM01home.html.

Boers, T., Olijkhoek, V., Van der Voort, C., Heijne, N., & Hidma, M. (2004). *Code 1. Basisleergang Nederlands voor anderstaligen*. Utrecht/Zutphen, Netherlands: ThiemeMeulenhoff.

Boersma, P., & Weenink, D. (2005). PRAAT. Retrieved from http://www.praat.org.

Bossers, B. (2010). Woordenschat (Vocabulary). In B. Bossers, F. Kuiken & A. Vermeer (Eds.), *Handboek Nederlands als tweede taal in het volwassenenonderwijs*. (Handbook Dutch as a second language in adult education) (pp. 166–207). Bussum, Netherlands: Coutinho.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Coste, D., Courtillon, J., Ferenczi, V., Martins-Baltar, M., & Papo, E. (1987). *Un niveau seuil*. Paris: Didier.

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

Council of Europe (2003). *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*. Pilot version. Strasbourg: Council of Europe.

Council of Europe (2009). *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*. Final version. Strasbourg: Council of Europe.

Dávid, G. (2007). Investigating the performance of alternative types of grammar items. *Language Testing*, *24*, 65–97.

Davidson, F., & Fulcher, G. (2007). The Common European Framework of Reference (CEFR) and the design of language tests: A matter of effect. *Language Teaching*, *40*, 231–241.

De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. H. (forthcoming, 2011).The effect of task complexity on native and non-native speakers' functional adequacy, aspects of fluency, and lexical diversity. In A. Housen, F. Kuiken & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency investigating complexity, accuracy and fluency in SLA*. Amsterdam, Netherlands: Benjamins.

De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. H. (forthcoming, 2012). Facets of speaking proficiency. *Studies in Second Language Acquisition*.

Decoo, W. (2011). *Systemization in foreign language teaching: monitoring content progression*. New York, NY: Routledge.

Engelen, R. J. H., & Eggen, T. J. H. M. (1993). Equivaleren. In T. J. H. M. Eggen & P. F. Sanders (Eds.), *Psychometrie in de praktijk* (pp. 309–348). Arnhem: Cito Instituut voor Toetsontwikkeling.

Fouly, K. A., Bachman, L. F., & Cziko, G. A. (1990). The divisibility of language competence: A confirmatory approach. *Language Learning*, *40*, 1–21.

Figueras, N., North, B., Takala, S., Verhelst, N., & Van Avermaet, P. (2005). Relating Examinations to the Common European Framework: A Manual. *Language Testing*, *22*, 261–279.

Harley, B., Cummins, J., Swain, M., & Allen, P. (1990). The nature of language proficiency. In B. Harley, P. Allen, J. Cummins & M. Swain (Eds.), *The development of second language proficiency* (pp. 7–25). Cambridge: Cambridge University Press.

Hulstijn, J. H. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal*, *91*, 663–667.

Hulstijn, J. H. (2011). Language proficiency in native and nonnative speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly*, *8*(3), 229–249.

Hulstijn, J. H., Alderson, J. C., & Schoonen, R. (2010). Developmental stages in second-language acquisition and levels of second-language proficiency: Are there links between them? In I. Bartning, M. Martin & I. Vedder (Eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research* (pp. 11–20). Eurosla Monographs Series, 1. Retrieved from http://eurosla.org/monographs/EM01/EM01home.html.

Jones, N., & Saville, N. (2009). European language policy: Assessment, learning, and the CEFR. *Annual Review of Applied Linguistics*, 29, 51–63.

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, *16*, 33–51.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol, UK: Multilingual Matters.

Milton, J. (2010). The development of vocabulary breadth across the CEFR levels. In I. Bartning, M. Martin & I. Vedder (Eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research* (pp. 211–232). Eurosla Monographs Series, 1. Retrieved from http://eurosla.org/monographs/EM01/EM01home.html.

Milton, J., & Alexiou, N. (2009). Vocabulary size and the Common European Framework of Reference for Languages. In B. Richards, H. M. Daller, D. D. Malvern, P. Meara, J. Milton & J. Treffers-Daller (Eds.), *Vocabulary studies in first and second language acquisition* (pp. 194–211). Basingstoke, UK: Palgrave.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge, UK: Cambridge University Press.

Nederlandse, Taalunie (2006). *Gemeenschappelijk Europees referentiekader voor moderne vreemde talen: Leren, onderwijs, beoordelen*. [No place.] Nederlandse Taalunie, Council of Europe.

Nederlandse Taalunie (2004). Corpus of spoken Dutch. Retrieved from http://lands.let.ru.nl/cgn/ehome.htm.

Read, J. (2000). *Assessing vocabulary*. Cambridge, UK: Cambridge University Press.

Shiotsu, T., & Weir, C. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Language Testing*, *24*, 99–128.

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, *36*(2), 139–152.

Tannenbaum, R. J., & Wylie, E. C. (2005). *Mapping English language proficiency test scores onto the Common European Framework*. Research Report 80. Princeton, NJ: Educational Testing Services.

Van Ek, J. A., & Trim, J. L. M. (1991). *Threshold 1990*. Strasbourg: Council of Europe.

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