Functional adequacy in L2 writing: Towards a new rating scale

Kuiken, F.; Vedder, I.

DOI
10.1177/0265532216663991

Publication date
2017

Document Version
Final published version

Published in
Language Testing

Article

LANGUAGE TESTING

Functional adequacy in L2 writing: Towards a new rating scale

Language Testing
2017, Vol. 34(3) 321–336

DOI: 10.1177/0265532216663991
journals.sagepub.com/home/ltj
SAGE

Folkert Kuiken and Ineke Vedder
University of Amsterdam, Netherlands

Abstract

The importance of functional adequacy as an essential component of L2 proficiency has been observed by several authors (Pallotti, 2009; De Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2012a, b). The rationale underlying the present study is that the assessment of writing proficiency in L2 is not fully possible without taking into account the functional dimension of L2 production. In the paper a rating scale for functional adequacy is proposed, containing four dimensions: (1) content, (2) task requirements, (3) comprehensibility, and (4) coherence and cohesion. The scale is an adaptation of the global rating scale of functional adequacy, which in an earlier study was carried out with expert raters (Kuiken, Vedder, & Gilabert, 2010; Kuiken & Vedder, 2014). The new rating scale for functional adequacy was tested out by a group of non-expert raters, who assessed the functional adequacy of a corpus of argumentative texts written by native and non-native writers of Dutch and Italian. The results showed that functional adequacy in L2 writing can be reliably measured by a rating scale comprising four different subscales.

Keywords

Assessing L2 writing, functional adequacy, (non-)expert raters, rating scale, rating scale level descriptors

The notion of language proficiency, as presented in the CEFR (Common European Framework of Reference for Languages; Council of Europe, 2001), can be defined both in terms of functions, domains and roles language users can handle ('can-do-statements'; *what*), and in terms of the quality of language proficiency (*how well*; Hulstijn, 2007; Hulstijn, Alderson, & Schoonen, 2010). Whereas the majority of studies in second language acquisition (SLA) conducted so far have been concerned with the linguistic

Corresponding author:
Folkert Kuiken, University of Amsterdam, Dutch Studies, Spuistraat 134, 1012 VB Amsterdam, Netherlands.
Email: f.kuiken@uva.nl

dimensions of L2 performance, operationalized as complexity, accuracy, and fluency (CAF), few studies report on the functional adequacy of L2 output. Furthermore, contrary to CAF research where general measures for assessing complexity, accuracy and fluency have often been employed (see, for an overview, Housen, Kuiken, & Vedder, 2012), general measures to rate the functional adequacy of L2 performance are lacking.

In this paper, we will present a new rating scale of functional adequacy (FA) in L2 writing. The rationale underlying the present study is (1) the assumption that the assessment of linguistic performance in L2 is not possible without taking into account the functional dimension of L2 production; (2) the necessity to assess FA as a separate dimension from CAF; (3) the view of FA as a multifaceted construct.

The outline of the paper is as follows: in the next section, the various definitions in the literature of FA and the way in which FA is usually assessed are discussed. We then propose our definition of FA as a multifaceted, task-related construct. In the third section, the relationship between FA and CAF is discussed. In the fourth section, we present an overview of studies which investigate the functional dimension of L2 production. In the fifth section, the new rating scale is presented, which was tested on both Dutch and Italian texts. The results are reported in the sixth section, together with a discussion of the applicability of the scale in the final section.

FA: Definition of the construct

The importance of FA as an essential component of L2 proficiency has been observed by authors such as Pallotti (2009), Kuiken et al. (2010), and De Jong et al. (2012a, b). However, whereas in language teaching and testing practice both the functional and the linguistic dimension of L2 performance are often independently assessed, in SLA research FA is often overlooked.

A possible reason for the paucity of studies in which the relationship between FA and the linguistic complexity, accuracy and fluency of L2 production is investigated, may be owing to the absence of a clear-cut definition of the construct. Until now, there has been no unanimity in the literature as to how FA is to be defined or assessed and by which features it is determined (Iwashita, Brown, McNamara, & O'Hagan, 2008). In recent studies in which FA is taken into account, the definition of FA varies greatly and various terms have been employed to designate the construct, such as communicative adequacy (Pallotti, 2009; Kuiken et al., 2010), communicative competence (McNamara & Roever, 2007), communicative effectiveness (Bridgeman, Powers, Stone, & Mollaun, 2012; Sato, 2012), intercultural competence (Hismanoglu, 2011), or communicative functionality (Fragai, 2001, 2003).

McNamara and Roever (2007) consider FA within a socio-pragmatic perspective, emphasizing the importance of the assessment of FA in L2 teaching. In other studies FA is defined on the basis of successful task performance (Pallotti, 2009; Alanen, Huhta, & Tarnanen, 2010; Kuiken et al., 2010), success of transfer of information (Upshur & Turner, 1995), or presented in an intercultural perspective (Hismanoglu, 2011). Relating linguistic and socio-pragmatic factors, Fragai (2001, 2003) distinguishes a number of features determining FA, such as adequate use of syntax and lexicon, and the ability to select appropriate linguistic forms for different communicative contexts. In the study by Knoch (2009,

2011), finally, FA is considered within a discourse-analytic perspective and defined in terms of coherence and cohesion of text (for a discussion of definitions of FA viewed from different theoretical angles, see also Leung, 2005, and Leung & Lewkowicz, 2012).

The definition of FA underlying the present study is that of FA viewed as a task-related, interpersonal construct, involving two participants (the writer A and the reader B). We thus consider FA in terms of successful task fulfilment (Alanen et al., 2010; Kuiken et al., 2010; De Jong et al., 2012a, b; Kuiken & Vedder, 2014). The focus in this definition is on the specific task to be carried out by A, and on the reception of the message by B. What is considered 'adequate' by the receiver B will depend on the particular language task (e.g. a letter to the student office, a short note to a house mate, an email to the teacher) to be performed by A. Given this relationship between successful task completion and adequacy of L2 output, we prefer to speak of *functional* rather than *communicative* adequacy. In terms of the conversational maxims of Grice (1975), the adequacy of the message transmitted by the writer A is understood and interpreted by the reader B with respect to the *quantity, relation, manner*, and *quality* of the message by A (see the 'Towards a new rating scale' section).

Functional adequacy and CAF

The relationship between the functional dimension and the linguistic dimension of L2 performance (e.g. complexity, accuracy, fluency) is not straightforward and to some extent problematic. Some utterances may score high on CAF parameters, but still be inadequate in relation to the task to be performed (Hulstijn, De Jong, Steinel, & Florijn, 2011). According to Pallotti (2009), FA can be viewed both as a separate dimension, independent from CAF, and as a way of interpreting it. In the first case, if FA is considered as an independent performance descriptor, it represents the extent to which a learner's performance is successful in achieving the task goal efficiently. The second way in which to consider FA in CAF research is in the interpretation of the CAF dimensions themselves. The assumption underlying this second interpretation of FA is that the more fluent, accurate or complex the learner's output is, the better it is from a functional perspective. As pointed out by Ortega (2003) and Pallotti (2009), many studies seem to take for granted that higher levels of CAF are to be positively evaluated and that lower levels of CAF depend on limitations in language-processing capacities. Both statements are questionable for the following reasons.

With regard to fluency, there appears to be an ideal rate of syllables per second, beyond which comprehension is affected (Kormos & Dénes, 2004). Also, pauses and dysfluencies can be communicatively functional in a particular situation, or may depend on the speaking style or personality traits of the speaker. Concerning accuracy, the mere fact that an L2 utterance is accurate with respect to L1 standard norms does not necessarily imply that it is adequate and pragmatically appropriate; conversely, non-standard, 'incorrect' forms may be pragmatically more adequate, as shown by Sanell (2007). In Sanell's study, Swedish advanced L2 learners of French tended to avoid the correct standard form of French negation (e.g. *je ne vais pas*), in spoken French, preferring instead the colloquial non-standard form (*je vais pas*), also used by native speakers.

Complexity is probably the most difficult dimension in relation to FA (Pallotti, 2009). 'More complex' does not necessarily mean 'better,' since a higher or lower complexity rate may be determined by personal and stylistic choices, rather than being an index of higher or lower FA (Ortega, 2003). Using overly complex but not ungrammatical sentences may be the result of L1 transfer, as was the case in the academic writing in English of learners with Spanish as L1 (Neff, Dafouz, Diez, Prieto, & Chaudron, 2004). Furthermore, a decrease of complexity in oral or written L2 production can sometimes be interpreted as a sign of higher proficiency, as demonstrated by Pallotti (2009). In the study, native and non-native speakers of Italian had to perform different speaking tasks: a film retelling and a service phone call to a shopkeeper or travel agency. The non-native students tended to produce long and complex syntactic units in both tasks, but sounded pragmatically unnatural in the phone task compared to the native speakers, who decomposed their utterances into syntactically simpler and shorter units. Over time, however, the L2 learners approached native speakers' behaviour by increasing their syntactic complexity in the retelling task and decreasing it in phone calls (Pallotti & Ferrari, 2008; Pallotti, 2009).

These studies by Ortega (2003), Pallotti (2009), and Pallotti and Ferrari (2008) demonstrate that the relationship between the development of FA and CAF in L2 is not unproblematic. 'Optimal' complexity levels may vary depending on what is considered functionally adequate in relation to a particular language task. Progress in L2 proficiency may thus include syntactic growth, but also entails the development of discourse and sociolinguistic repertoires in L2, to be adapted appropriately by the language learner to particular communication demands and language tasks (Ortega, 2003).

The role of functional adequacy in L2 research

A few studies, conducted within the framework of TBLT (task-based language teaching), also have taken into account the functional dimension of L2 performance. The majority of these studies, as mentioned in the previous section, have addressed FA in oral communication, rather than in written discourse. However, despite clear differences between the two modes, the methodology, and the type of rating scales that were employed for assessing FA in oral discourse may partly be applicable also to writing. Although in the studies discussed here, different terms have been used to designate the functional aspects of L2 output, in this section and in the previous sections we adopt the term FA, as pointed out earlier.

A study by De Jong et al. (2012a, b) aimed to decompose the construct of language proficiency, investigating to what extent learner differences in linguistic sub-skills are related to individual differences in successfully conveying information through speaking. This study was performed on the spoken production of L2 students of Dutch. A panel of non-linguists rated the communicative performance of 181 L2 learners through different tasks, on the basis of both the amount of information, its relevance, the setting (formal/informal), the discourse type (descriptive/persuasive), and the intelligibility of the response. The rating scales comprised six levels. The results showed that, next to intonation skills, vocabulary turned out to be the variable more tightly associated with FA.

A study by Sato (2012) on FA in oral proficiency investigated the relative contribution of linguistic criteria and FA to overall intuitive judgments of L2 speaking proficiency in an oral test of English for academic purposes (EAP). In the study, 9 expert raters (L2 teachers of English) had to assess the oral performance of 30 Japanese undergraduate students with respect to FA, grammatical accuracy, fluency, vocabulary, and pronunciation, on a scale from one to five. Next to these ratings they had to write an open-ended comment for each recording. The results showed that raters tended to assess FA in a way consistent with overall intuitive judgments of speaking proficiency. Fluency was perceived as the most salient among linguistic criteria. Other linguistic criteria, such as accuracy, vocabulary, and pronunciation, contributed less to raters' intuitive judgments.

In a study by Bridgeman et al. (2012), ratings of FA of speech samples produced by L2 learners of English assigned by experienced TOEFL raters, undergraduate students, and by an automated scoring system (SpeechRater), were compared. FA was operationalized in terms of the amount of effort expended to understand the speech samples. FA was evaluated both by rating scales and by the ability of the raters to answer multiple-choice questions that could be answered only if the spoken response was understood. The results of the study demonstrated that correlations between the FA scores of the expert raters with those assigned by undergraduate raters were substantially higher than correlations with the scores provided by SpeechRater.

One of the few studies on FA in L2 writing was conducted by Kuiken et al. (2010), who investigated in their CALC (communicative adequacy and linguistic complexity) study the relationship in L2 writing between FA and linguistic complexity for Dutch, Italian, and Spanish. The analyses were carried out on the basis of written data from 32 learners of Dutch, 39 learners of Italian, and 23 learners of Spanish, with a proficiency level ranging from A2 to B1. All participants were subjected to two writing tasks, comprising a short argumentative text. As a benchmark the same writing tasks were administered to a group of 17 native speakers of Dutch, 18 of Italian, and 10 of Spanish. All texts in L2 and L1 were rated by expert raters (all of them both native speakers and L2 teachers of the target language) on both FA and linguistic complexity on six-point global Likert scales.

Interrater reliability in the CALC-study was found to be acceptable to good (Cronbach's α ranged between .70 and .90). In order to get more insight into the reasons why a particular rating score had been assigned and which linguistic features were associated by the raters with a particular scale level, two panel discussions were organized, one for Dutch and one for Italian. The main finding of the study was that FA and linguistic complexity in L2 appeared to develop at the same pace. The correlation between the two dimensions, however, tended to be higher for high-level learners compared to low-level learners. This may show that more advanced L2 learners, who do not have to devote all their cognitive resources to linguistic form, may have more attentional and memory resources available to deal with the functional aspects of the text. An implication of the study was that it is important to have FA assessed not only by expert but also by non-expert raters, since experienced raters/L2 teachers may be 'biased' and more lenient or more strict, compared to 'naive' native speakers (Kuiken & Vedder, 2014).

The studies discussed here demonstrate the necessity to develop a valid, reliable, and theory-based rating scale, in order to assess FA as a crucial dimension of L2 proficiency

(Fulcher, 1987; McNamara, 2002; Knoch, 2011). This raises a number of questions concerning the criteria and scale descriptors of a rating scale of FA in L2 writing, the underlying theoretical premises and the relationship to CAF. Another issue concerns the applicability of the scale also for non-expert raters. In the following section, on the basis of the results of the studies on the role of FA reviewed so far, we propose a new rating scale of FA in written L2 performance, inspired by Grice's (1975) maxims, comprising four different subscales and six scale levels, which can be used by both expert and non-expert raters. The focus of the study at hand is on the reliability of the scale.

Towards a new rating scale

The new rating scale of FA is an adaptation of the one that we used in the CALC-study (Kuiken et al., 2010; Kuiken & Vedder, 2014), which was based on two main sources: the general descriptors provided by the CEFR (Council of Europe, 2001) and the scale descriptors employed for developing the rating scales of De Jong et al. (2012a, b). The requirements of the proposed rating scale of FA are the following: (1) deconstruction of relevant components of FA; (2) independence of FA descriptors from linguistic descriptors in terms of CAF; (3) 'objective' and 'countable' scale descriptors; (4) applicability both for expert and non-expert raters; and (5) the possibility of using the scale in both L2 and L1. The scale (a six-point Likert scale) is inspired by the conversational maxims of Grice (1975), focusing on the quantity, relevance, manner and quality of the message of the writer A to be transmitted to the interlocutor B.

The new rating scale of FA, defined in terms of successful task completion of A in conveying a message to B and in relation to the conversational maxims of Grice, thus comprises the following four scale dimensions (see Appendix A).

Content: is the number of information units (i.e. 'ideas' or 'concepts') provided in the text adequate and relevant?

This dimension takes into account (1) the adequacy of the number and type of information units in the text, and (2) their consistency and relevance, independently from the specific requirements of the language task to be carried out (cf. Grice, 1975: maxims of quantity and relation). For instance, in the case where learners have to write an argumentative text on the importance of physical exercise, information units could concern ideas expressed as 'increase of obesity,' 'risk for a heart attack,' 'a sitting lifestyle,' and so on.

Task requirements: have the requirements of the task been fulfilled successfully with respect to, for example, text genre, register, and speech act?

This dimension focuses on the specific instructions and requirements of the task to be completed and the adequacy of the message transmitted by the writer A to the reader B (cf. Grice, 1975: maxim of quality).

Comprehensibility: how much effort is required of the reader B in order to understand the text purpose and ideas expressed by the writer A?

This dimension takes into account the extent to which the message of the writer A turns out to be adequate and comprehensible for a particular reader B (cf. Grice, 1975: maxim of manner; see also Bridgeman et al., 2012; De Jong et al., 2012a, b).

Coherence and cohesion: is the text written by the writer A coherent and cohesive?

This dimension focuses on the adequacy of the message of the writer A for the interlocutor B in terms of the occurrence of cohesive ties (presence or absence of deictic elements, anaphoric devices, and strategies) conjunction use, coherence breaks, number of repetitions (Knoch 2007, 2009, 2011).

Testing out the new rating scale

In order to test out the reliability of the new rating scale of FA, the data collected in the CALC project, discussed in the section 'The role of functional adequacy in L2 research' (Kuiken et al., 2010; Kuiken & Vedder, 2014), were presented to a group of non-expert raters, who were asked to judge FA in the two writing texts assigned to the L2 learners. As was also the case in Kuiken and Vedder (2014), the two writing tasks produced by the L2 learners of Dutch and Italian were employed, while the data related to Spanish L2 were omitted for practical reasons. The texts written by the L1 writers of Dutch and Italian were used as a benchmark.

Research questions

The central research question discussed in the paper is as follows: how do non-expert raters judge the functional adequacy of argumentative texts written by L2 learners of Dutch and Italian by means of a six-point Likert rating scale containing four dimensions of FA?

The following questions will be addressed:

1. How do the judgments of raters correlate with each other in terms of the interrater reliability scores on the four dimensions of FA?
2. How are the judgments of raters on the four dimensions of FA correlated?
3. How do the judgments of raters of FA in the texts written by L2 learners correlate with the judgments of the texts written by L1 writers?
4. How do the judgments of raters of FA in the two writing tasks written by the same participants correlate?

Tasks and texts

The texts used in the present study were collected from 32 learners of Dutch L2 and 39 learners of Italian L2, with a proficiency level ranging from A2 to B1 (see the previous section, 'The role of functional adequacy in L2 research'). All participants were subjected

Table 1. Intra-class correlations among raters for L2 and L1 responses.

Dimension	Dutch	Italian
Content	.841**	.838**
Task requirements	.824**	.725**
Comprehensibility	.940**	.901**
Coherence and cohesion	.860**	.867**

*$p < .05$, **$p < .01$, ***$p < .001$.

to two writing tasks, prompting the participants to write two short argumentative texts. In the first task, learners were required to make a decision about which of three non-governmental organizations to choose as a candidate for receiving a grant, whereas in the second task they had to decide which topic they would like to see published as the main article in the monthly magazine of their favourite newspaper (for a more detailed description, see Kuiken et al., 2010). As a benchmark the same writing tasks were administered to a group of 17 native speakers of Dutch and 18 native speakers of Italian.

Raters and scales

Both the L2 and L1 texts were rated on a six-point Likert scale by non-expert raters on four dimensions of FA: content, task requirements, comprehensibility, and coherence/cohesion (see the 'Towards a new rating scale' section). Both for Dutch and for Italian there were four raters, with either Dutch or Italian as their mother tongue, respectively. All of them were university students of approximately the same age as the L2 and L1 writers involved in the study. The raters did not have any specific experience in judging written texts and can therefore be categorized as being non-expert. In order to get accustomed to working with the four dimensions of FA, two training sessions were organized. During these sessions, the aims and use of the rating scale were explained and raters were trained by means of models and practice materials. In the next section, the results of the use of the rating scale of FA by these non-expert raters will be presented.

Results

We will start this section by reporting the interrater reliability scores obtained by the raters of Dutch and Italian while judging the L2 and L1 texts. We will then present the correlations between the four dimensions of FA, and the similarities between the judgments of the texts written by the L2 learners and those produced by the native writers. Finally, we will report on the extent to which raters concur in their judgments of the two texts written by the same participants.

Interrater reliability scores

In order to assess the interrater reliability scores, intra-class correlations were calculated (for L2 and L1 combined) among the four raters of Dutch and the four raters of Italian (Table 1). Interrater reliability was high across all raters, ranging between .725 (for task

Table 2. Pearson's product–moment correlations between dimensions of FA for L2 and L1 participants).

Dimension	Dutch			Italian		
	Task requirements	Comprehensibility	Coherence and cohesion	Task requirements	Comprehensibility	Coherence and cohesion
Content	.848**	.814**	.880**	.710**	.844**	.877**
Task requirements		.694*	.851*		.544**	.559**
Comprehensibility			.873**			.938**

*$p < .05$, **$p < .01$, ***$p < .001$.

requirements in Italian) and .940 (for comprehensibility in Dutch). These correlations were all statistically significant ($p < .01$ for all correlation pairs).

Correlations between dimensions of functional adequacy

To find out the extent to which the judgments of the raters on the four separate dimensions of functional adequacy agree with each other, Pearson's product–moment correlations were calculated (see Table 2). It turns out that all correlations are significant and differ from moderate to strong. The relation between comprehensibility and coherence/cohesion is particularly strong (.873 for Dutch and .938 for Italian). There also appears to be a strong correlation with other dimensions of FA (varying from .710 with task requirements for Italian to .880 with coherence/cohesion for Dutch). Lower values are obtained for correlations between task requirements and other dimensions of FA, in particular for Italian (.544 with comprehensibility and .559 with coherence/cohesion).

Rater judgments of L2 vs. L1 texts

Not surprisingly, raters assign higher scores on all dimensions of FA to L1 writers than to L2 learners (see Table 3). Scores for Dutch L2 range from 2.84 for coherence/cohesion to 3.04 for task requirements; for Dutch L1 scores vary from 3.96 for content to 4.79 for comprehensibility. For Italian a similar difference between L2 and L1 writers emerges: scores for L2 range from 3.54 for coherence/cohesion to 4.32 for task requirements; L1 scores range from 4.94 for coherence/cohesion to 5.39 for comprehensibility. These numbers also demonstrate that on the whole, the L2 writers of Italian obtain higher scores than the L2 writers of Dutch. As standard deviations are also somewhat higher for Italian (.43 to .82) than for Dutch (.18 to .60), there appears to be greater variety between the writers of Italian than between the writers of Dutch.

From a paired samples *t*-test, one can conclude that the observed differences between writers in L2 and L1 are significant for all four dimensions of FA, except for that of task requirements for Italian, which is not significant at the corrected alpha value of .01 (see Table 4); a Bonferroni correction was used to correct for the use of multiple paired-sample *t*-tests, by dividing the alpha value by the number of tests (.05/4 = .0125).

Table 3. Descriptive statistics for Dutch and Italian scores divided by L2 and L1 participant groups.

Dimension	L2–L1	Dutch			Italian		
		N	Mean	*SD*	*N*	Mean	*SD*
Content	L2	32	2.92	.46	39	3.87	.70
	L1	17	3.96	.55	18	5.03	.63
Task requirements	L2	32	3.04	.59	39	4.32	.69
	L1	17	4.04	.60	18	4.74	.82
Comprehensibility	L2	32	3.09	.47	39	3.80	.73
	L1	17	4.79	.18	18	5.39	.43
Coherence and cohesion	L2	32	2.84	.45	39	3.54	.61
	L1	17	4.16	.47	18	4.94	.49

Table 4. Paired samples *t*-test between L2 and L1.

Dimension	Dutch				Italian			
	df	*t*	Mean diff	*p*	*df*	*t*	Mean diff.	*p*
Content	47	−7.010	−1.041	<.001***	55	−5.961	−1.156	<.001***
Task requirements	47	−5.588	−.998	<.001***	55	−2.017	−.419	.049
Comprehensibility	47	−14.370	−1.693	<.001***	55	−8.583	−1.588	.001**
Coherence and cohesion	47	−9.584	−1.318	<.001***	55	−8.491	−1.399	.001**

*p < .05, **p < .01, ***p < .001; significant at alpha = .0125.

Correlations between texts produced by the same writer

Finally, we investigated whether raters differed in their judgments of the two texts written by the same participant. On the basis of the calculation of Pearson's correlation coefficients one can conclude that the correlations between the two texts (combined for L2 and L1) are in all cases high for the four dimensions of functional adequacy in terms of internal reliability consistency, indicating that raters judge both texts more or less in the same way (see Table 5). The lowest correlation is obtained on task requirements for Italian (.455), the highest on comprehensibility for Dutch (.877). All correlations are significant ($p < .001$).

Conclusion and discussion

In this study we have defined FA as successful task fulfilment, in terms of the conversational maxims of Grice (1975) of quantity, relation, manner and quality. On the basis of the maxims of Grice we developed a six-point scale of FA comprising four different scale dimensions. The requirements are that the descriptors should be objective and countable, independent from CAF measures (i.e. complexity, accuracy and fluency),

Table 5. Pearson's product–moment correlations between task 1 and task 2.

Dimension	Dutch			Italian		
	N	Correlation	Significance	*N*	Correlation	Significance
Content	49	.623***	<.001	57	.607***	<.001
Task requirements	49	.704***	<.001	57	.455***	<.001
Comprehensibility	49	.877***	<.001	57	.766***	<.001
Coherence and cohesion	49	.719***	<.001	57	.802***	<.001

*$p < .05$, **$p < .01$, ***$p < .001$.

and that application of the scale should be possible in L2 and L1, both by expert and non-expert raters.

The results of the study show that FA can reliably be measured by a rating scale containing four different subscales. Content was assessed by considering the number and relevance of information units. Content, as we operationalized it, is distinct from task requirements, which focus on the specific task to be carried out in terms of genre, register, and speech act. For raters it appeared to be fairly easy to distinguish the two constructs. Also, comprehensibility, operationalized in terms of the amount of effort and 're-reading' required of the reader in order to understand the text, could be assessed in a countable way. This also was true for coherence and cohesion, which could be assessed in a quantifiable way, by considering the presence or absence of cohesive and anaphoric devices, strategies, coherence breaks, and repetitions.

The intra-class correlations among the four raters varied from acceptable to excellent. The correlation between the subscale for content and the other dimensions of functional adequacy was high, which also was the case between comprehensibility and coherence/cohesion. Lower correlations were found, not surprisingly, between task requirements and comprehensibility, and between task requirements and coherence/cohesion for Italian. It is, indeed, perfectly possible to write a comprehensible and coherent text that is, however, unsuccessful in terms of specific task requirements. The scales could also be employed for rating the texts of both L2 and L1 writers, although raters found it easier to differentiate between L2 learners compared to L1 writers, who all performed at the upper end of the scale. The differences in FA scores that were observed between Dutch L2 and Italian L2 are likely to be owing to differences in the type of learners: the Italian L2 writers were first-year students of Italian with Dutch as their L1, whereas the Dutch L2 learners, who were preparing for university, all came from different countries and varied in their mother tongues. Correlation coefficients of the ratings on both writing tasks produced by the same participants were good and in all cases significant, implying that raters were generally stable in their judgments of the two texts.

All in all, the study demonstrates that FA, as a fundamental component of L2 proficiency, can be reliably assessed. Based on the findings, we propose that a multidimensional construct as FA should be assessed by means of a rating scale comprising four dimensions. Being conceptually different, it is important to distinguish these dimensions, despite the significant correlations among them. As shown by this study, the rating scale

is applicable for non-expert raters, and it turned out to be possible for non-expert raters to become familiar with the scales in just two training sessions. The new rating scale of FA thus appears to be a reliable and efficient tool for assessing written learner production. The scale also may have pedagogical advantages, allowing teachers to provide L2 writers with specific feedback and focused comments on the writing skills of the learners on the four different dimensions of FA.

A number of issues need to be investigated in further research. In this study, we have focused on the findings of the groups of L2 writers of Dutch and Italian as a whole, without taking into account differences among individual learners and raters. It may be necessary to look more in detail at individual differences among L2 writers and to focus also on the judgments and perceptions of FA of each individual rater. Finally, it should be kept in mind that in our study the new rating scale of FA has been tested out for adult, highly educated L2 learners of Dutch and Italian, who were subjected to one type of task, in one particular mode (writing an argumentative text). The applicability of the scale should, however, also be tested out with other groups of participants (e.g. learners with a low level of education, adolescents), with different types of speaking and writing tasks, and for languages other than Dutch and Italian. A better insight into the development of FA in L2 writing, finally, also may shed more light on related issues, such as the role of L1 transfer, or the acquisition of 'diagnostic' textual features, distinguishing different levels of L2 proficiency (Hulstijn et al., 2010).

Acknowledgements

The authors wish to thank all student raters for their evaluation of the data of Dutch and Italian. We particularly thank Chiara Sale (University of Cagliari) for her assistance and Saskia Nijman (University of Groningen) for the statistical analysis of the data. We also thank Gabriele Pallotti (University of Modena and Reggio Emilia), and our colleagues of CASLA (Cognitive Approaches in Second Languages Acquisition; University of Amsterdam) for their comments.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Alanen, R., Huhta, A., & Tarnanen, M. (2010). Designing and assessing L2 writing tasks across proficiency levels. In I. Bartning, M. Martin & I. Vedder (Eds.), *Communicative proficiency and linguistic development. Intersections between SLA and language testing research* (pp. 21–56). Eurosla Monographs Series 1.

Bridgeman, B., Powers, D., Stone, E., & Mollaun, P. (2012). TOEFL iBT speaking test scores as indicators of oral communicative language proficiency. *Language Testing*, *29*(1), 91–108.

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

De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012a). The effect of task complexity on functional adequacy, fluency and lexical diversity in speaking performances of native and non-native speakers. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 121–142). Amsterdam: John Benjamins.

De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012b). Facets of speaking proficiency. *Studies in Second Language Acquisition*, *34*(1), 5–34.

Fulcher, G. (1987). Tests of oral performance: The need for data-based criteria. *ELT Journal*, *41*(4), 28–91.

Fragai, E. (2001). La programmazione didattica: Il Glotto-Kit come strumento per valutare i livelli in entrata. In M. Barni & A. Villarini (Eds.), *La questione della lingua per gli immigrati stranieri. Insegnare, valutare e certificare l'Italiano L2* (pp.191–208). Milan: Franco Angeli Editore.

Fragai, E. (2003). Valutare la competenza linguistico-comunicativa in italiano L2: Il Glotto-Kit per bambini e adolescenti stranieri. *Didattica & Classe Plurilingue*, *7*, 1–5.

Ghout-Khenoune, L. (2012). The effects of task type on learners' use of communication strategies. *Procedia. Social and Behavioral Sciences*, *69*, 770–779.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Speech acts* (pp. 41–58). New York: Academic Press.

Hismanoglu, M. (2011). An investigation of ELT students' intercultural communicative competence in relation to linguistic proficiency, overseas experience and formal instruction. *International Journal of Intercultural Relations*, *35*(6), 805–817.

Housen, A., Kuiken, F., & Vedder, I. (Eds.) (2012). *Dimensions of L2 performance and proficiency. Complexity, accuracy and fluency in SLA*. Amsterdam: John Benjamins.

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., 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.

Hulstijn, J. H., De Jong, N. H., Steinel, M., & Florijn, A. F. (2011). 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.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, *29*(1), 24–49.

Knoch, U. (2007). 'Little coherence, considerable strain for reader': A comparison between two rating scales for the assessment of coherence. *Assessing Writing*, *12*(2), 108–128.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, *26*(2), 275.

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing*, *16*(2), 81–96.

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32*(2), 146–164.

Kuiken, F., & Vedder, I. (2014). Rating written performance: What do raters do and why? *Language Testing*, *31*(3), 329–348.

Kuiken, F., Vedder, I., & Gilabert, R. (2010). Communicative adequacy and linguistic complexity in L2 writing. In I. Bartning, M. Martin, & I. Vedder (Eds.), *Communicative proficiency*

and linguistic development: Intersections between SLA and language testing research (pp. 81–100). Eurosla Monographs Series 1.

Leung, C. (2005). Convivial communication: Recontextualizing communicative competence. *International Journal of Applied Linguistics*, *15*(2), 119–144.

Leung, C., & Lewkowicz, J. (2012). Language communication and communicative competence: A view from contemporary classrooms. *Language and Education*, *26*(6), 1–17.

McNamara, T. (2002). Discourse and assessment. *Annual Review of Applied Linguistics*, *22*, 221–242.

McNamara, T., & Roever, C. (2007). *Testing: The social dimension*. Malden, MA/Oxford, UK: Blackwell.

Neff, J., Dafouz, E., Diez, F., Prieto, R., & Chaudron, C. (2004). Contrastive discourse analysis: Argumentative texts in English and Spanish. In C. I. Moder & A. Martinovic-Zic (Eds.), *Discourse across languages and cultures* (pp. 267–283). Amsterdam: John Benjamins.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, *24*(4), 492–518.

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, *30*(4), 590–601.

Pallotti, G., & Ferrari, S. (2008). La variabilità situazionale dell'interlingua: Implicazioni per la ricerca acquisizionale e il testing linguistico. In G. Bernini, L. Spreafico, & A. Valentini (Eds), *Competenze lessicali e discorsive nell'acquisizione di lingue seconde* (pp. 437–461). Perugia: Guerra.

Sanell, A. (2007). *Parcours acquisitionnel de la négation et de quelques particules de portée en français L2*. Doctoral thesis, Stockholm University.

Sato, T. (2012). The contribution of test-takers' speech content to scores on an English oral proficiency test. *Language Testing*, *29*(2), 223–241.

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal*, *49*(1), 3–12.

Appendix A. Rating scale for functional adequacy.

Content: Is the number of information units provided in the text adequate and relevant?

1	2	3	4	5	6
The number of ideas is *not at all adequate* and insufficient and the ideas are unrelated to each other.	The number of ideas is *scarcely adequate* and the ideas lack consistency.	The number of ideas is *somewhat adequate*, even though they are not very consistent.	The number of ideas is *adequate* and they are sufficiently consistent.	The number of ideas is *very adequate* and they are very consistent to each other.	The number of ideas is *extremely adequate* and they are very consistent to each other.

Task requirements: Have the task requirements been fulfilled successfully (e.g. genre, speech acts, register)?

1	2	3	4	5	6
None of the questions and the requirements of the task have been answered.	*Some (less than half)* of the questions and the requirements of the task have been answered.	*Approximately half* of the questions and requirements of the task have been answered.	*Most (more than half)* of the questions and the requirements of the task have been answered.	*Almost all* the questions and the requirements of the task have been answered.	*All* the questions and the requirements of the task have been answered.

Comprehensibility: How much effort is required to understand text purpose and ideas?

1	2	3	4	5	6
The text is *not at all comprehensible*. Ideas and purposes are unclearly stated and the efforts of the reader to understand the text are ineffective.	The text is *scarcely comprehensible*. Its purposes are not clearly stated and the reader struggles to understand the ideas of the writer. The reader has to guess most of the ideas and purposes.	The text is *somewhat comprehensible*. Some sentences are hard to understand at a first reading. A second reading helps to clarify the purposes of the text and the ideas conveyed, but some doubts persist.	The text is *comprehensible*. Only a few sentences are unclear but are understood, without too much effort, after a second reading.	The text is *easily comprehensible* and reads smoothly. Comprehensibility is not an issue.	The text is *very easily comprehensible* and highly readable. The ideas and the purpose are clearly stated.

(Continued)

Appendix A. (Continued)

Coherence and cohesion: Is the text coherent and cohesive (e.g. cohesive devices, strategies)?

1	2	3	4	5	6
The text is *not at all coherent*. Unrelated progressions and coherence breaks are very common. The writer does not use any anaphoric device. The text is *not at all cohesive*. Connectives are hardly ever used and ideas are unrelated.	The text is *scarcely coherent*. The writer often uses unrelated progressions; when coherence is achieved, it is often done through repetitions. Only a few anaphoric devices are used. There are some coherence breaks. The text is *not very cohesive*. Ideas are not well linked by connectives, which are rarely used.	The text is *somewhat coherent*. Unrelated progressions and/or repetitions are frequent. More than two sentences in a row can have the same subject (even when the subject is understood). Some anaphoric devices are used. There can be a few coherence breaks. The text is *somewhat cohesive*. Some connectives are used, but they are mostly conjunctions.	The text is *coherent*. Unrelated progressions are somewhat rare, but the writer sometimes relies on repetitions to achieve coherence. A sufficient number of anaphoric devices is used. There may be some coherence breaks. The text is *cohesive*. The writer makes good use of connectives, sometimes not limiting this to conjunctions.	The text is *very coherent*: when the writer introduces a new topic, it is usually done by using connectives or connective phrases. Repetitions are very infrequent. Anaphoric devices are numerous. There are no coherence breaks. The text is *very cohesive* and ideas are well linked by adverbial and/or verbal connectives.	The writer ensures *extreme coherence* by integrating new ideas in the text with connectives or connective phrases. Anaphoric devices are used regularly. There are few incidences of unrelated progressions and no coherence breaks. The structure of the text is *extremely cohesive*, thanks to a skillful use of connectives (especially linking chunks, verbal constructions and adverbials), often used to describe relationships between ideas.