E-JournALL, EuroAmerican Journal of Applied Linguistics and Languages
Volume 7, Issue 2, October 2020, 9–27
ISSN 2376-905X
http://doi.org/10.21283/2376905X.12.175
www.e-journall.org

Assessing functional adequacy across tasks: A comparison of learners' and native speakers' written texts[1]

ELENA NUZZO
GIUSEPPE BOVE
Università Roma Tre

Received 17 February 2019; accepted after revisions 10 June 2020

ABSTRACT

EN This study aims to explore further the applicability of the six-point rating scale elaborated by Kuiken and Vedder (2017) to assess functional adequacy. According to the authors, functional adequacy (FA) is a multidimensional construct combining successful task completion and the effective transmission of a message from a speaker or writer to a hearer or reader. Kuiken and Vedder's scale has been mainly applied to written L2 output in previous research, whereas in this study it is tested on both L1 and L2 written texts, elicited by means of different tasks. The texts, produced by 20 non-native and 20 native speakers of Italian, were rated by seven non-expert raters, trained to use the scale. The results showed low levels of absolute inter-rater agreement and consistency, especially for L1 texts. On the other hand, our findings are more encouraging with regard to the reliability of the scale and its applicability across tasks. The results also revealed a strong correlation between FA scores and those obtained in a C-test used as an independent measure of general L2 proficiency.

Key words: FUNCTIONAL ADEQUACY, L1 ITALIAN, L2 ITALIAN, ASSESSMENT, WRITING TASK

ES El presente estudio tiene como objetivo profundizar en la aplicabilidad de la escala de calificación de seis puntos elaborada por Kuiken y Vedder (2017) para evaluar la adecuación funcional. Según los autores, la adecuación funcional (AF) es un constructo multidimensional que combina la finalización satisfactoria de una tarea y la transmisión efectiva de un mensaje de un hablante o escritor a un oyente o lector. En investigaciones anteriores, la escala de Kuiken y Vedder se ha aplicado principalmente a la producción escrita de L2, mientras que en este estudio se prueba en textos escritos tanto de L1 como de L2, obtenidos mediante diferentes tareas. Los textos, producidos por 20 hablantes no nativos y 20 hablantes nativos de italiano, fueron calificados por siete evaluadores no expertos que habían sido capacitados para emplear la escala. Los resultados muestran bajos niveles de concordancia y consistencia absolutas entre los evaluadores, especialmente en los textos de L1. Por otro lado, los hallazgos son más alentadores con respecto a la fiabilidad de la escala y su aplicabilidad en las tareas. Además, los resultados revelan una fuerte correlación entre las puntuaciones de AF y las obtenidas en una prueba C utilizada como medida independiente de la competencia general de L2.

Palabras clave: ADECUACIÓN FUNCIONAL, ITALIANO L1, ITALIANO L2, EVALUACIÓN, TAREAS DE ESCRITURA

IT Questo studio ha l'obiettivo di approfondire l'applicabilità della scala di valutazione a sei punti elaborata da Kuiken e Vedder (2017) per valutare l'adeguatezza funzionale. Secondo gli autori, l'adeguatezza funzionale (FA) è un costrutto multidimensionale che combina lo svolgimento positivo del compito e la trasmissione efficace di un messaggio da un parlante o scrittore a un ascoltatore o lettore. La scala di Kuiken e Vedder, che finora era stata applicata principalmente alla produzione scritta in L2, in questo studio è stata testata su elaborati scritti sia in L1 sia in L2 ottenuti a partire da consegne diverse. I testi, prodotti da 20 parlanti non nativi e 20 madrelingua italiani, sono stati valutati da sette valutatori non esperti, addestrati all'uso della scala. I risultati mostrano bassi livelli di accordo e coerenza assoluti tra i valutatori, soprattutto per i testi in L1. I risultati sono, invece, più incoraggianti per quanto riguarda l'affidabilità della scala e la sua applicabilità a varie attività. Inoltre, i risultati rivelano una forte correlazione tra i punteggi FA e quelli ottenuti da un C-test utilizzato come misura indipendente della competenza generale in L2.

Parole chiave: ADEGUATEZZA FUNZIONALE, ITALIANO L1, ITALIANO L2, VALUTAZIONE, COMPITI DI SCRITTURA

✉ **Elena Nuzzo,** Università Roma Tre
elena.nuzzo@uniroma3.it

[1] The authors worked together throughout the paper. E. Nuzzo wrote sections 1, 2.1, 2.2, 2.3, 4, and 5; G. Bove wrote sections 2.4 and 3.

The assessment of language proficiency is not fully possible without considering the functional dimension of language use, especially from the can-do perspective which informs the Common European Framework of Reference for Languages (Council of Europe, 2001). In order to assess the degree to which a linguistic performance is more or less successful in achieving the goals of a task, one needs to take into account its adequacy, and not only its quality in terms of complexity, accuracy and fluency (CAF measures, see Housen, Kuiken, & Vedder, 2012 for an overview). In fact, a speaker's production which scores high on CAF can be inadequate because, among other factors, the chosen topics are partially or totally irrelevant, the register inappropriate, or the sequence of ideas ineffective. In contrast, a linguistic performance may be effective and adequate even if it is neither complex nor accurate nor fluent (Pallotti, 2009, p. 596).

Kuiken and Vedder observed that few SLA studies report on the functional aspects of L2 proficiency, and that "contrary to CAF research where general measures for assessing complexity, accuracy and fluency have often been employed . . . general measures to rate the functional adequacy of L2 performance are lacking" (2017, p. 322). They therefore proposed a rating scale for what they called *functional adequacy* (FA), a construct characterized by a strong relationship between task fulfillment and the appropriateness of linguistic output. The new scale was successfully tested out with written and oral L2 argumentative texts. However, its applicability to L1 performances and to a variety of tasks is still underexplored (Kuiken & Vedder, 2018).

The present study aims to fill this gap. For the first time, the applicability of the scale is explored on both L1 and L2 output and on different types of written texts, with the same team of raters. The decision to investigate further the applicability of this scale with native and non-native speakers stems from the previous experience of one of the authors as pre-service and in-service teacher trainer, as well as teacher in courses of Italian for academic purposes. It seems that the development of rating instruments that can measure and describe functional aspects of language proficiency in both L1 and L2 would be helpful for the growing number of teachers who—in schools and universities—work simultaneously with non-native and native speakers and are traditionally more inclined to give emphasis to morphosyntactic accuracy rather than to functional aspects of language use.

The paper is organized as follows: Firstly, we introduce the construct of functional adequacy and the scale, and then review the studies in which it has been applied. Next, we present the aims of this study and the methodology applied. The two following sections are devoted to the presentation and discussion of results. In the last section, some pedagogical implications are discussed together with concluding remarks.

1. Functional adequacy: Construct and rating scale

In recent years, several SLA studies have devoted attention to the functional dimension of L2 production. Although they all have in common a focus on the functional aspects of language proficiency, definitions and labels vary. For example, Kuiken, Vedder, and Gilabert (2010) and Kuiken and Vedder (2014) used the term *communicative adequacy* (see also Pallotti, 2009; Révész, Ekiert, & Torgersen, 2016), meaning a task-related, dynamic and interpersonal construct which focuses on both the specific task being carried out by the speaker, and the way the message is received by the interlocutor. DeJong et al. (2012a, 2012b) referred to the successful conveyance of messages through speaking with the expression *functional adequacy*. Bridgeman et al. (2012) and Sato (2012) adopted the label *communicative effectiveness*, meaning success of information transfer. The notion of *communicative competence* in McNamara and Roever (2007) focuses on the sociopragmatic appropriateness of the linguistic means within the specific context of use.

According to Kuiken and Vedder, the lack of unanimity as to how the functional dimension of linguistic output is to be defined or assessed could explain why this dimension is still often overlooked in SLA research, as well as why general measures to rate this aspect of L2 performance are lacking (2017, p. 322). In an attempt to address this gap, Kuiken and Vedder (2017) outlined a systematic proposal for defining and measuring the functional aspects of language proficiency. They chose the label *functional adequacy* (FA) and defined the construct as "successful task completion of A in conveying a message to B and in relation to the conversational maxims of Grice" (1975, p. 326), thus focusing on the specific task carried out by the speaker or writer and on the reception of the message by the addressee. According to this interpretation, FA is therefore intended as a task-related and interpersonal construct, characterized by a strong relationship between successful task fulfillment and the appropriateness and effectiveness of linguistic output.

A six-point Likert rating scale (Appendix A) was developed by Kuiken and Vedder (2017) as a tool to measure FA. The scale, which is an adaptation of the one they used in previous studies (Kuiken et al., 2010;

Kuiken & Vedder, 2014), comprises four subscales, corresponding to the four dimensions of FA identified by the authors, namely content, task requirements, comprehensibility, coherence and cohesion.

The content dimension focuses both on the adequacy of the number and type of ideas provided in the text, and on their consistency and relevance to the general topic, regardless of the specific requirements of the task to be completed. It refers to Grice's (1975) maxims of quantity and relation. The dimension of task-requirements, related to Grice's (1975) maxim of quality, takes into account the specific instructions of the task to be carried out, particularly in terms of register, genre, and speech acts. Comprehensibility deals with the effort which is needed for reader or listener B to understand the ideas expressed by the writer or speaker A (maxim of manner; Grice, 1975). The coherence and cohesion dimension takes into account the use of textual cohesion mechanisms, such as for example anaphoric devices, deictics, connectives, and the occurrence of coherence breaks (maxim of manner; Grice, 1975).

According to the authors, the rating scale they propose should meet the following criteria: (i) "deconstruction of relevant components of functional adequacy"; (ii) "independence of FA descriptors from linguistic descriptors in terms of CAF"; (iii) "'objective' and 'countable' scale descriptors"; (iv) "applicability both for expert and non-expert raters"; and (v) applicability to both L2 and L1 (Kuiken & Vedder, 2017, p. 326). It is this last criterion that we specifically address in the present paper.

1.1. The application of the rating scale

The scale was first successfully tested with L2 (Dutch and Italian) written argumentative texts assessed by non-expert raters (Kuiken & Vedder, 2017; see also Vedder, 2016). Given that in a previous study (Kuiken & Vedder, 2014) experienced raters appeared to be "biased" and stricter or more lenient compared to "naive" native speakers, the authors opted for non-expert raters. Kuiken and Vedder (2017) found good intraclass correlations among raters, and high correlations between the four subscales. More specifically, a high correlation was found between content and the other dimensions, and between comprehensibility and coherence/cohesion. Yet, correlations between task requirements and comprehensibility were lower. In the retrospective focus group, the raters did not report any particular difficulties in using the scales, which seemed thus applicable also for non-expert raters. In sum, the study confirmed that FA can be assessed as a multidimensional construct comprising four dimensions, and the authors concluded that the "new rating scale of FA thus appears to be a reliable and efficient tool for assessing written learner production" (p. 332).

After this first empirical investigation, the new rating scale was applied in a number of studies with similar—albeit not identical—characteristics, and with different learner populations. Most of these studies used L2 written texts (Del Bono, 2017; Faone & Pagliara, 2017; Orrù, 2019; Pagliara, 2017), but some attempts were made with L2 spoken output (Kuiken & Vedder, 2018). All of the studies involved non-expert raters trained to use the scale before the rating sessions. The results were not always consistent across studies, as will be shown in the following paragraphs.

Faone and Pagliara (2017) tested the scale with instruction texts written by Chinese learners of Italian. Their data showed good inter-rater reliability (Cronbach's α ≥ 0.89) but low levels of inter-rater agreement (intraclass correlations values less than 0.57). Correlations between the subscales are from moderate to high (most values greater than 0.74; only in two cases involving the comprehensibility subscale correlated with content and task requirement subscales; values are 0.50 and 0.57 respectively). Similar results were found in a replication of the study with narrative texts written by the same Chinese learners (Pagliara, 2017). Unfortunately, in both studies retrospective discussions with the raters were lacking.

Del Bono (2017) and Orrù (2019) applied the scale to narrative, instruction, and argumentative texts produced by Dutch and Hungarian learners of Italian. The tasks were the same used in the present study. On the whole, Del Bono's (2017) results are similar to those reported in Faone and Pagliara (2017) and in Pagliara (2017), whereas Orrù (2019) found good levels of inter-rater reliability (all greater than 0.8) and agreement (all greater than 0.77). Both studies also explored the correlation between FA scores and an independent measure of general proficiency: the same C-test used in the present study. A high correlation coefficient (higher than 0.78) was found by Del Bono (2017) for narrative and instruction tasks, less so (greater than 0.73 for each task) by Orrù (2019). On the other hand, the two authors agree in reporting that their non-expert raters encountered some difficulties in using the scales, particularly as far as the coherence and cohesion dimension is concerned.

In Kuiken and Vedder (2018) the scale was adapted for spoken monological texts, and a comparison was proposed between written and oral data. Non-expert raters assessed oral and written argumentative texts produced by two groups of university students of L2 Dutch and L2 Italian. According to the results obtained, the scale appears to be a reliable tool for assessing the functional adequacy of both written and spoken L2 production.

As for the use of the scale with L1 output, fewer investigations have been completed, and the results appear rather mixed. In the previously mentioned study carried out to test the scale (Kuiken & Vedder, 2017), L1 samples were also used. The authors concluded that the scale could be employed "for rating the texts of both L2 and L1 writers" (p. 331). Nevertheless, they reported that raters found it easier to discriminate among L2 learners than among L1 informants, whose scores tended to cluster at the upper end of the scale.

On the other hand, in a study where the scale was applied only to native speakers' (oral and written) data, it turned out that raters used the whole scale and discriminated among the performances. However, inter-rater reliability scores—measured using the index r_{WG} of inter-rater agreement within groups proposed by James et al. (1984)—were below the 0.7 threshold for some of the dimensions, particularly for written data (Cortés Velásquez & Nuzzo, 2017).

To sum up, these studies show a great variability in their results, providing a much less clear and optimistic picture compared to the first application of the scale by Kuiken and Vedder (2017). It appears that differences in research design result in significant variation of findings. In particular, the few attempts made to use the scale with L1 texts lead to rather contradictory results. It seems therefore that more research is needed to investigate further the applicability of the scale in a variety of contexts, in order to gain a clearer picture of the potential utility of this tool.

2. Goal and methods

The general aim of this study is to examine from three different perspectives the applicability of the rating scale developed by Kuiken and Vedder (2017) for the assessment of FA in written texts produced by native and non-native speakers of Italian performing three different tasks.

The first (and most important) perspective is the applicability of the rating scale as an instrument to assess the position of an individual performance with respect to an absolute threshold (e.g., the minimum level of the scale to pass an exam or the level under which a remedial class is required), in a criterion-referenced interpretation of the rating scale (Graham, Milanowsky, & Miller, 2012, p. 6). From this perspective, inter-rater absolute agreement (IRA) of the rating scale should be analyzed, that is the degree to which two or more evaluators (raters) using the scale give the same rating to an identical target (e.g., student, teacher, etc.). High levels of IRA are required for the application of the scale in a criterion-referenced perspective. Consequently, the first research question addressed is:

> RQ1: How do raters' judgments compare in terms of inter-rater absolute agreement scores on the four dimensions of FA in narrative, instruction, and argumentative texts written by native and non-native speakers of Italian?

A rating scale with low levels of IRA can still be considered from a second perspective, concerning its applicability as an instrument to assess the relative position of an individual performance with respect to performances of other individuals in a group (or to the mean or median performance of the group), in a norm-referenced view of the rating scale. This can be useful, for instance, in ranking participants in any kind of competition (e.g., for a scholarship, a job position, etc.). In this case, it is the consistency between evaluators in the ordering or relative standing of performance ratings (inter-rater consistency) that should be analyzed, regardless of the absolute values of evaluator's rating (Graham, Milanowsky, & Miller, 2012, p. 5). High levels of inter-rater consistency are required for the application of the scale in a norm-referenced perspective. Consequently, the second research question asks:

> RQ2: How do raters' judgements compare in terms of inter-rater consistency scores on the four dimensions of FA in narrative, instruction, and argumentative texts written by native and non-native speakers of Italian?

Finally, the third and last perspective from which a rating scale is considered in scientific research is its applicability as an instrument for measuring a latent construct (i.e., its reliability). In this case researchers should use all the information available from all evaluators (including discrepant ratings), attempting to create a summary score (for each dimension and for the global FA scale in our study) for each participant (Stemler & Tsai, 2008, p. 42). What is relevant is that most of the variance of the defined summary score (e.g., the raters' mean rating or some other function of the raters' rating) is shared between the ratings, because this gives an indication that the evaluators are rating a common construct (Stemler & Tsai, 2008, p. 43). It is not important that the evaluators provide exactly the same ratings for each participant, nor that their ratings are ordered consistently, because each evaluator is seen as providing also some unique information that is useful in estimating the summary score of the participant. High levels of reliability of the rating scale are required for the application of the scale as an instrument for measuring a latent construct. Since the FA scale is composed of four dimensions, the following three research questions are addressed:

> RQ3: What are the reliability levels of the four dimensions of FA in narrative, instruction, and argumentative texts written by native and non-native speakers of Italian?
>
> RQ4: What are the reliability levels of the global FA scale within each type of text written by native and non-native speakers of Italian?
>
> RQ5: How do the raters' evaluations of the texts written by native and non-native speakers of Italian correlate with general levels of proficiency as measured by a C-test?

As we saw in the previous section, the FA scale was applied to L1 data in only two studies, both relying solely on argumentative texts, and the results were conflicting. For this reason, the present study can be considered exploratory in nature and no hypotheses are proposed.

2.1. Informants and data collection

To calculate the sample sizes for the present study, designs based on intraclass correlation coefficients reported in Doros and Lew (2010) and Gwet (2014, pp. 249-251) were taken into account. Following these approaches, it was possible to determine sample sizes (number of writers and raters) based on the expected value of the intraclass correlation and on a chosen width of the 95% confidence interval (i.e., twice the size of the maximum error). In order to have a maximum error 0.15 (confidence interval length equal to 0.3), and assuming an expected intraclass correlation approximately equal to 0.7, it was required to select at least twenty writers and a number of raters between five and seven.

Twenty native speakers of Italian and twenty speakers of L2 Italian participated in the study as writers. The native speakers were students of foreign languages at Roma Tre University. Their age ranged from 19 to 22, with a mean of 20.15. Fifteen of the non-native speakers were students attending a course of Italian linguistics at the University of Amsterdam; their L1 was Dutch, except for one whose native language was English. The remaining five non-native speakers were students of foreign languages at Roma Tre University. They had different L1s, namely Chinese, French, Hungarian, Romanian, and Spanish. The mean age in the non-native speakers group was 26.89, with the majority ranging from 19 to 25 and seven informants aged between 28 and 55. Their overall language proficiency was measured by a C-test. Ten learners scored between 48 and 66, nine between 73 and 88, and one had a score of 96[2].

Each writer produced three written texts on the basis of three tasks (see further details in the next section and instructions in Appendix B). The participants were given 90 minutes to write the texts and to complete the C-test (see the next section for a description of the test), and were not allowed to use the dictionary. The information about the writers and the data collection is summarized in Table 1.

Table 1
Synthesis of informants' data

Informants	n	Mean Age	Tasks	Texts
L1 writers	20	20.15	3	60
L2 writers (various L1s)	20	26.89	3	60
Total	**40**	**23.52**	**3**	**120**

[2] The C-test was administered to the native speakers as well. The majority scored between 97 and 100, whereas three of them had lower scores (90, 94, and 96). These data were not considered in the analysis.

2.2. The three tasks and the C-test

The tasks, which were developed at the University of Amsterdam for two studies involving Spanish L2 learners of English[3] and Hungarian L2 learners of Italian (Orrù, 2019) and then translated into Italian (Del Bono, 2017), were of three types, namely a narrative task, an instruction task, and a decision-making task. In the first task the informants were required to write a short story in which they described an episode that occurred during a study trip abroad. The second task required them to write a message to a couple who would hypothetically rent their place for a week and give instructions about the apartment. In the third task they had to write an email to a university office and explain which housing option they preferred, and why, among three different residence types during a study abroad program.

The C-test (the same used in Kuiken et al., 2010) consisted of five short texts in which half the letters of every other word had been replaced by blanks, for a total of 100 incomplete words. The informants had to reconstruct the words on the basis of contextual clues.

2.3. Raters and rating procedure

All the texts produced by L1 and L2 writers (120 texts in total, see Table 1) were assessed by seven native speakers of Italian on the six-point Likert scale of FA elaborated by Kuiken and Vedder (2017). The raters were female university students of about the same age as the informants involved in the study, attending the same faculty, or department, at the same level (MA). The raters did not have previous experience in assessing written texts and can therefore be considered non-expert raters[4]. Following Kuiken and Vedder (2017), two training sessions were organized to familiarize the raters with the four dimensions of FA. During the training sessions, the researchers illustrated the aims of the FA scale and how to use it, and the raters were trained through models and hands-on practice.

The raters worked individually on the texts, which were assigned to them without any details about the authors. This means that the raters did not know whether they were rating a text written by a native or a non-native speaker. However, they might have found out the status of some of the writers by noting the presence (or absence) of morphosyntactic errors, which native speakers would not generally make.

After the rating sessions, a panel discussion was organized, during which the raters were asked to verbalize the reasons behind their judgments, any difficulties in using the scale, and the strategies they used when assessing the texts.

2.4. Data analysis

Descriptive statistics (mean and standard deviation) were computed for each dimension to compare L1 and L2 groups and highlight trends across the three tasks. The differences in mean scores were tested for statistical significance using standard independent samples *t* tests. Besides a basic description of our sample, this also allows a first comparison with previous studies, in particular with Kuiken and Vedder (2017).

Inter-rater agreement can be measured in different ways according to the aim of the analysis and the criteria for the selection of observations (e.g., Gwet, 2014; McGraw & Wong, 1996). To answer research questions RQ1 and RQ2, absolute agreement and consistency were considered, respectively. When analyzing absolute agreement, we were interested in establishing to what extent raters' evaluations were close to an equality relation (e.g., in the case of only two raters, if the two sets of ratings are represented by x and y the relation of interest is x = y). In the case of consistency, the relation was relaxed (e.g., additive: y = x + a or linear: y = bx + a). Consistency is thus a different concept from absolute agreement because it allows raters to use different ranges and/or scale units. In this study, both aspects have been considered.

Due to the characteristics of our research design, the most appropriate measure for inter-rater absolute agreement was the intraclass correlation coefficient ICC(A,1) presented in McGraw and Wong (1996), which we computed among the seven raters for each dimension in each task. Inter-rater consistency for each dimension-task combination was assessed using intraclass correlation coefficient ICC(C,1) (McGraw & Wong, 1996) and the average inter-rater Pearson's correlation (i.e., the average of the twenty-one Pearson's correlations obtained comparing each pair of raters). The first coefficient relates to additive consistency, whereas the average inter-rater Pearson's correlation reflects linear consistency.

[3] This study was still being conducted while we were writing the present paper.

[4] It was decided to use non-expert raters to allow for better comparability with previous studies on the scale.

For the reader interested in the psychometric aspects of our study, we would like to remark that intraclass correlation coefficients can be considered from the point of view of generalizability theory (e.g., Fan & Sun, 2014). More specifically, the intraclass correlation coefficient ICC(A,1) can be considered equivalent to generalizability coefficient ϕ in a one-facet G-study (e.g., Fan & Sun, 2014, p. 56).

To answer research question RQ3, reliability analysis for each dimension of the FA scale was conducted in each task by measuring Cronbach's alpha, both for the total group of participants and for L1 group and L2 group separately. A satisfactory reliability level ($\alpha > 0.7$) supports the possibility to summarize the construct represented by the dimension with the average scores provided for each writer. In our case, satisfactory levels of reliability were obtained in the L2 group but not in the L1 group, so it seemed worth exploring further the latent structure of the FA scale in the three tasks only for the L2 group. For each L2 participant, the scores assigned by the seven raters were averaged for each dimension. The three correlation matrices (one for each task) between the average scores of the four dimensions of FA were computed and analyzed by principal component analysis to compare the factor structures obtained for each task and to answer research question RQ4. Given the small number of variables (the four dimensions in each of the three matrices), in each task the analysis of the factor structure was aimed to check the unidimensionality hypothesis, under which a single latent construct of FA (the first principal component) is sufficient to account for the relations among dimensions (a similar approach is considered in Stemler & Tsai, 2008, p. 42-43). The percentage of variance explained by the first principal component and levels of the factor loadings of the four dimensions were taken into consideration (explained variance equal or higher 70% and factor loadings equal or higher than 0.7 for all dimensions were considered to support the unidimensionality hypothesis). Support to the unidimensionality hypothesis allows us to measure the latent construct by the computation of a global score of FA in each task that can be used for correlational studies involving other linguistic constructs (e.g., CAF measures) or socio-cultural features of respondents.

Finally, to answer RQ5, Pearson's correlation coefficients were calculated between the global FA scores in each task and the C-test scores, to analyze the association between the FA scale and an indicator of general proficiency (are the two measures linearly independent, or are some aspects of functional adequacy associated with proficiency? If so, to what extent?).

All the analyses presented in this and the following section were conducted by the statistical software package IBM SPSS (release 24).

3. Results

3.1. Comparison of L2 vs. L1 rater judgments

Descriptive statistics for L2 and L1 texts, and for the whole sample, are reported in Table 2. For each writer and each dimension, mean scores of the ratings provided by the seven raters were computed. Then for each task, the group mean scores were computed for L1, L2, and the whole sample (L1-L2). Mean scores are clustered at the upper end of the scale (mainly ranging between 4 and 6). The small range of mean scores is clearly visible in Figure 1, where a curve is associated with each task. The curves have similar trends because the raters assigned higher scores to L1 informants on all dimensions across the three tasks. This is in line with the results in Kuiken and Vedder's (2017) study, though they found lower upper and lower range limits, as well as less dispersion than the present study. Task 2 has higher mean scores compared to the other two tasks. L2 scores show more relative dispersion than L1 scores, when compared to the corresponding mean scores (i.e., dispersion measured by coefficients of variation).

Table 2
Descriptive statistics for L2, L1, and total scores (Task 1 – Narrative, Task 2 – Instruction, Task 3 – Decision Making)

		TASK														
DIMENSION	**GROUP**	**Task 1**					**Task 2**					**Task 3**				
	L2-L1	N	Mean	Min	Max	SD	N	Mean	Min	Max	SD	N	Mean	Min	Max	SD
Content	L2	20	3.89	2.29	5.29	1.27	20	4.50	3.00	5.71	1.17	20	4.47	3.14	5.29	1.06
	L1	20	4.99	4.00	5.86	1.02	20	5.01	2.14	5.86	1.18	20	4.99	3.86	5.71	1.01
	L1-L2	40	4.44	2.29	5.86	1.27	40	4.75	2.14	5.86	1.20	40	4.73	3.14	5.71	1.07
Task requirements	L2	20	4.16	2.29	5.14	1.23	20	4.51	2.57	5.57	1.10	20	4.09	2.71	5.29	1.09
	L1	20	5.11	4.43	5.71	0.77	20	5.04	3.29	5.43	0.87	20	4.59	3.29	5.29	0.97
	L1-L2	40	4.64	2.29	5.71	1.13	40	4.77	2.57	5.57	1.03	40	4.34	2.71	5.29	1.06
Comprehensibility	L2	20	4.41	1.43	5.57	1.19	20	4.56	2.00	5.71	1.07	20	4.52	2.71	5.57	0.96
	L1	20	5.67	4.86	6.00	0.62	20	5.79	5.43	6.00	0.44	20	5.66	5.00	5.86	0.57
	L1-L2	40	4.87	1.43	6.00	1.23	40	4.68	2.00	6.00	1.23	40	4.90	2.71	5.86	1.11
Coherence cohesion	L2	20	3.83	2.43	5.14	1.04	20	3.80	2.57	4.86	0.93	20	3.76	2.57	4.71	0.97
	L1	20	4.80	4.00	5.29	0.93	20	4.74	3.43	5.29	1.06	20	4.76	3.71	5.57	0.95
	L1-L2	40	4.31	2.43	5.29	1.10	40	4.27	2.57	5.29	1.10	40	4.26	2.57	5.57	1.08

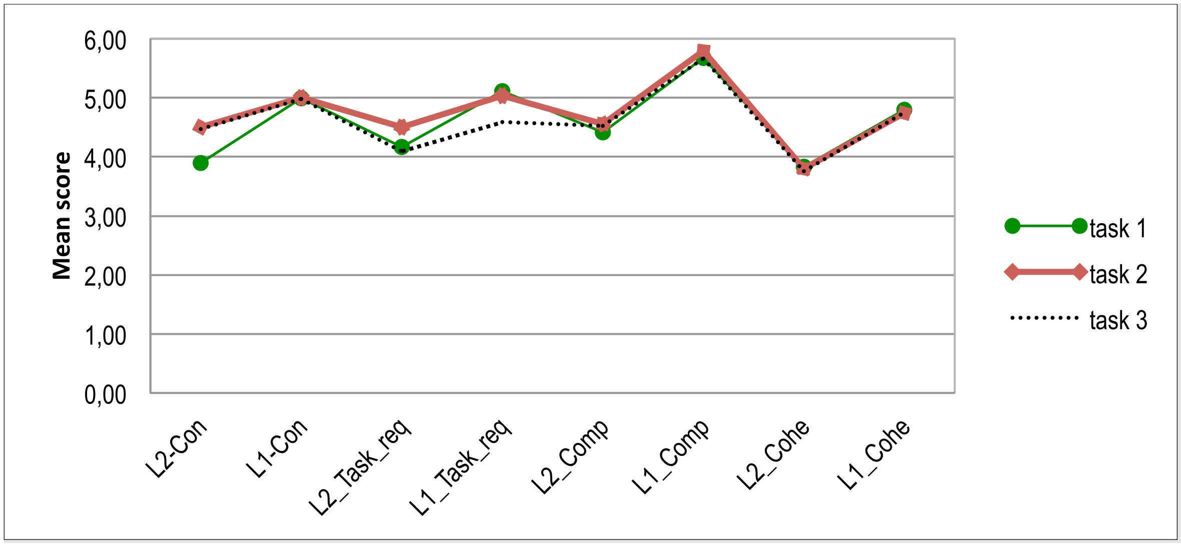

Figure 1. Task mean scores

Independent samples *t* tests on the difference between mean scores of L2 and L1 texts gave the results reported in Table 3. As one would expect, the texts written by the native speakers were rated significantly higher than those produced by the non-native speakers. The only exception is the content dimension in Task 2, where the difference between the mean scores of the two groups of writers is not significant.

Table 3

Independent sample t test between L2 and L1 groups

	Task 1		
DIMENSION	**Mean diff. (L2-L1)**	**t**	**df**
Content	-1.10	-4.78***	38
Task requirements	-.94	-4.79***	38
Comprehensibility	-1.26	-5.39***	38
Coherence and cohesion	-0.97	-5.86***	38
		Task 2	
Content	-.51	-1.93	38
Task requirements	-.53	-2.58*	38
Comprehensibility	-1.23	-6.31***	38
Coherence and cohesion	-0.94	-6.20***	38
		task 3	
Content	-.51	-2.81**	38
Task requirements	-.49	-2.42*	38
Comprehensibility	-1.14	-6.69***	38
Coherence and cohesion	-1.00	-6.16***	38

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

3.2. Inter-rater absolute agreement (RQ1)

Absolute agreement between raters' judgments was measured by intraclass correlation coefficients (ICC(A,1)), as explained in the data analysis section. Results are reported in Table 4. Considering all the tasks, ICC values on the four dimensions for the total L2-L1 sample range from 0.27 to 0.68, showing a low level of agreement between raters. When compared to Tasks 1 and 2, Task 3 had lower intraclass correlation coefficient values, with the exception of the coherence and cohesion dimension.

If we analyze L2 and L1 groups separately, intraclass correlation coefficient ranges from 0.24 to 0.63 for the former, and from 0.02 to 0.43 for the latter. L2 values are generally higher than L1 values, in particular for the comprehensibility dimension, where raters' judgments reach the highest intraclass correlation coefficient in the L2 group and the lowest in the L1 group. The dispersion of the ratings assigned on each dimension (computing the corresponding box-plots) was explored, and it was found that the small range of scores assigned by the raters, especially when judging L1 texts, may explain these very low values of intraclass correlations. We report the box-plots of raters' judgments on the comprehensibility dimension in Task 2 (Figure 2). The L1 panel shows a very low degree of variability, as most of the raters used only levels 5 and 6. A similar picture can be found for the comprehensibility dimension in the other two tasks, and, to a lesser extent, for the coherence and cohesion dimension in the three tasks. In these situations, participant average scores (i.e., the averages of the scores assigned by all the raters to each participant) show low variability, and this restriction of the between-writer variance determines low intraclass correlation coefficients.

Table 4
Intraclass correlation coefficients (ICC(A,1))

		TASK					
		Task 1		Task 2		Task 3	
DIMENSION	**GROUP**	**N**	**ICC**	**N**	**ICC**	**N**	**ICC**
Content	L2	20	0.33	20	0.42	20	0.27
	L1	20	0.29	20	0.43	20	0.20
	L1-L2	40	0.43	40	0.44	40	0.27
Task requirements	L2	20	0.33	20	0.38	20	0.41
	L1	20	0.15	20	0.29	20	0.19
	L1-L2	40	0.40	40	0.38	40	0.34
Comprehensibility	L2	20	0.63	20	0.59	20	0.50
	L1	20	0.14	20	0.02	20	0.06
	L1-L2	40	0.67	40	0.68	40	0.59
Coherence and cohesion	L2	20	0.30	20	0.27	20	0.24
	L1	20	0.09	20	0.11	20	0.17
	L1-L2	40	0.35	40	0.30	40	0.36

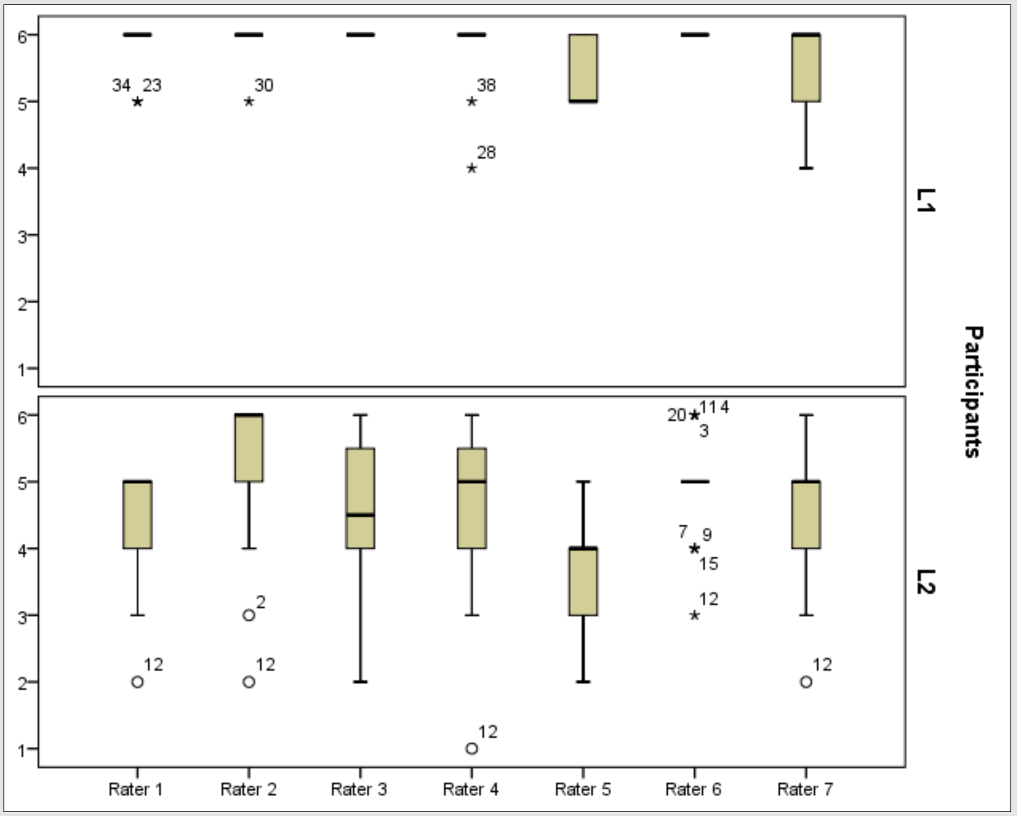

Figure 2. Box-plots of raters judgments on the comprehensibility dimension in Task 2

3.3. Inter-rater consistency (RQ2)

The level of consistency between raters' judgments was measured by ICC(C,1) and Pearson correlations. Results are reported in Table 5 for the whole group of participants. In particular, for each dimension-task combination the (7 × 7) inter-rater Pearson correlation matrix was analyzed. For each matrix, the average inter-rater Pearson's correlation was computed, along with the minimum correlation, the maximum correlation and percentage of correlations greater or equal to 0.7. Intraclass correlation values and average inter-rater correlations totally agree, indicating a low level of consistency between raters' judgments. The only exception is the comprehensibility dimension, which shows a moderate consistency in Tasks 1 and 2. Separate analyses of inter-rater consistency for the L1 and L2 groups (not reported here) confirmed the results obtained for the total group L1-L2.

Table 5
Intraclass correlation coefficients (ICC(C,1)) and average inter-rater Pearson's correlations for the total group L2-L1

DIMENSION	**ICC**	**Ave(r)**	**Min(r)**	**Max(r)**	**r ≥0.7**
			Task 1		
Content	0.51	0.51	-0.26	0.76	9.5%
Task requirements	0.48	0.46	0.05	0.79	9.5%
Comprehensibility	0.71	0.72	0.60	0.85	81.0%
Coherence and cohesion	0.47	0.46	-0.03	0.70	4.8%
			Task 2		
Content	0.55	0.57	0.35	0.83	4.8%
Task requirements	0.46	0.47	0.27	0.68	0.0%
Comprehensibility	0.76	0.77	0.68	0.88	95.2%
Coherence and cohesion	0.47	0.48	0.04	0.70	4.8%
			Task 3		
Content	0.39	0.38	0.04	0.66	0.0%
Task requirements	0.47	0.45	0.16	0.73	4.8%
Comprehensibility	0.62	0.63	0.52	0.75	19.0%
Coherence and cohesion	0.45	0.45	0.11	0.69	4.8%

3.4. FA scale reliability analysis (RQ3-RQ4)

In order to explore the latent structure of the FA scale, as a first step Cronbach's alpha was measured for each dimension-task combination. Results are reported in Table 6. In the three tasks, the alpha values of the four dimensions for the whole L2-L1 sample range from 0.82 to 0.96. Higher values were obtained for the comprehensibility dimension (average α = 0.94) compared to the other dimensions (average α = 0.86 each). Task 3 has slightly lower values (average α = 0.86) compared to the other two tasks (average α = 0.89 each).

When we analyze the L2 and L1 groups separately, things change. The alpha values range from 0.73 to 0.94 for the L2 group, and from 0.16 to 0.90 for the L1 group. In most cases L2 alpha values are higher than L1 alpha values, in particular for the comprehensibility dimension, where alpha values reach the lowest values for the L1 group in tasks 2 (α = 0.16) and 3 (α = 0.33). Results concerning alpha values showed a good reliability level for the L2 group for all dimensions and tasks. For the L1 group a good reliability level was obtained for the content dimension in each task, for the task requirement dimension only in Tasks 2 and 3, and for the coherence and cohesion dimension only in Task 3. Unsatisfactory reliability levels were obtained for the comprehensibility dimension in the L1 group in all tasks. This is a consequence of the strong range restriction of the scores, mostly concentrated on levels 5 and 6 of the scale. This range restriction results in low inter-rater correlations and low alpha values.

Table 6
Cronbach's alphas for dimensions and tasks

		TASK					
DIMENSION	**GROUP**	**Task 1**		**Task 2**		**Task 3**	
		N	α	N	α	N	α
Content	L2	20	0.83	20	0.88	20	0.82
	L1	20	0.80	20	0.90	20	0.75
	L1-L2	40	0.88	40	0.89	40	0.82
Task requirements	L2	20	0.83	20	0.87	20	0.89
	L1	20	0.66	20	0.79	20	0.75
	L1-L2	40	0.87	40	0.85	40	0.86
Comprehensibility	L2	20	0.93	20	0.94	20	0.89
	L1	20	0.60	20	0.16	20	0.33
	L1-L2	40	0.95	40	0.96	40	0.92
Coherence and cohesion	L2	20	0.80	20	0.77	20	0.73
	L1	20	0.63	20	0.69	20	0.77
	L1-L2	40	0.86	40	0.86	40	0.85

In summary, according to Cronbach's alpha values, it seemed worth exploring further the latent structure of the FA scale in the three tasks only for the L2 group. To this aim, for each L2 participant the scores assigned by the seven raters were averaged for each FA dimension. Pearson's correlation coefficients were calculated between the dimensions for L2 participants, and results are reported in Table 7. Correlation coefficients ranged from moderate to high in tasks 1 and 2, whereas lower values were found in Task 3 for the comprehensibility dimension.

The structure of the three correlation matrices was also studied through principal component analyses for the three tasks. In each task only the first principal component was selected (by the criterion of the eigenvalues greater than 1). The first principal component explained 81% of variance in the first task, 80% in the second task and 69% in the third task (Table 8). In each task factor loadings on the first principal component were equal or higher than 0.8 (most values higher than 0.9, and only a low value 0.66 for the comprehensibility dimension in the third task), supporting the unidimensionality hypothesis for the FA scale across the tasks. This result allows the computation of a global score of FA in each task, that could be used in correlational studies involving other linguistic constructs (e.g., CAF proficiency) or socio-cultural features of respondents.

Table 7
Pearson's product-moment correlations between dimensions for L2 group

DIMENSION	**Task requirements**	**Comprehensibility**	**Coherence and cohesion**
		Task 1	
Content	0.78**	0.69**	0.86**
Task requirements		0.75**	0.75**
Comprehensibility			0.68**
		Task 2	
Content	0.86**	0.63**	0.73**
Task requirements		0.66**	0.77**
Comprehensibility			0.77**
		Task 3	
Content	0.76**	0.54*	0.77**
Task requirements		0.18	0.65**
Comprehensibility			0.57**

*p < .05, **p < .01

Table 8

Loadings and explained variance of the first principal component for each task

	First Principal Component		
DIMENSION	**Task 1**	**Task 2**	**Task 3**
Content	0.93	0.90	0.94
Task requirements	0.91	0.92	0.80
Comprehensibility	0.86	0.85	0.66
Coherence and cohesion	0.91	0.91	0.91
Explained variance	81.3%	80.2%	69.4%

3.5. Correlation between the global FA scores and the C-test scores (RQ5)

The three global FA scores are significantly correlated with C-test scores, particularly for the instruction task (0.87). Moderately high correlations were found for the narrative (0.70) and the decision-making (0.61) tasks. This indicates that the levels of the global FA scale and the levels of proficiency as measured by the C-test are associated, especially for the instruction task (e.g., writers with high levels of the proficiency according to the C-test usually obtain high global ratings on the FA scale). These results seem to indicate that proficiency as measured by the C-test is not independent from FA, and that a high degree of association can be observed at least with some aspects of the FA scale across tasks.

4. Discussion

The general purpose of this study was to explore further the applicability of the six-point rating scale for FA recently proposed by Kuiken and Vedder (2017). In particular, we aimed to put the scale to the test with both L2 and L1 written texts elicited by means of three different tasks.

Low levels of inter-rater agreement and consistency, especially for the native speakers' group, suggest that the scale is not equally applicable to L1 and L2 written output by non-expert raters. This seems to be in agreement with the findings of Kuiken and Vedder (2014), who reported that it was difficult for their (expert) raters to judge L1 and L2 production at the same time, by means of the same communicative adequacy rating scale (a previous version of the FA scale used here). The authors therefore observed that the rating scales developed for the assessment of L2 written output are not always suitable for the assessment of L1 writing (Kuiken & Vedder, 2014, p. 333), possibly because L2 writing has its own particular characteristics. However, when the new FA scale was tested, the findings revealed that it could be employed "for rating the texts of both L2 and L1 writers" (Kuiken & Vedder, 2017, p. 331). The only problem was that raters found it more difficult to discriminate among native speakers as opposed to non-native speakers.

If it is true that our findings do not confirm the applicability of the FA scale to both L1 and L2 writing, the low values of inter-rater agreement found in our data may be explained as a consequence of the small range of scores provided by the raters, particularly when judging L1 texts. In fact, little variation of subject scores results in low intraclass correlation coefficients. The low variability in scores might have also caused the low degree of consistency found between raters' judgments.

As far as the reliability of the FA scale is concerned, positive results were obtained. The good correlation levels between the four dimensions support the unidimensionality hypothesis for the scale across the three tasks, meaning that the four dimensions represented by the sub-scales are part of the single latent construct of FA. This is particularly true for the whole L2-L1 sample and the L2 group alone, whereas in the native speakers' group alone lower reliability levels were found, especially on the comprehensibility dimension. Again, this is a consequence of the range restriction of the scores—mostly concentrated on levels 5 and 6 of the scale for the L1 group—which results in low inter-rater correlations. The applicability of the scale as a measurement instrument to different types of tasks seems to be supported by our results, at least for the L2 data.

High correlation values were achieved between global FA scores and C-test scores in the L2 group, for all tasks but particularly for the instruction one. Global language proficiency as measured by the C-test seems thus to be somehow associated to FA, in that those who obtain high ratings on the former usually score high on the latter as well. This result, which is in line with Del Bono's (2017) findings but not with Orrù's (2018, forthcoming), suggests that FA is part of the general proficiency of a speaker or writer, even though it can be assessed independently from descriptors of linguistic performance in terms of CAF.

During the panel discussion that followed the assessment sessions, our non-expert raters reported several difficulties in using the scales, with some of the subscales described as more challenging than the others. Also, according to their comments, they sometimes felt uncomfortable separating functional adequacy from linguistic accuracy, possibly because in their experience as students the assessment of written texts is inherently associated with the latter.

As in Del Bono (2017) and Orrù (2019), the coherence and cohesion scale was described as the most difficult to use, mainly due to the fact that the two aspects are not always as closely linked as the scale descriptors would suggest. For example, the raters reported they had found texts that were totally coherent even though no connectives or anaphoric devices were used. However, this perceived difficulty with one of the sub-scales seems to not have affected negatively the use of the scale as a whole, since in the analysis the four dimensions represented by the sub-scales proved to be part of the single latent construct of FA.

Furthermore, the raters claimed that they found the scale not equally applicable to the three tasks. More specifically, they declared they had experienced difficulties in using the scale with narrative texts (Task 1), which they considered more creative in nature and therefore hardly evaluable in terms of set criteria and determined standards.

On the whole, our results are not as anticipated, particularly with regard to the applicability of the FA scale to L1 texts. However, they confirm that the scale has a good potential for assessing the FA of different types of written texts, given that its reliability as an instrument to measure a latent construct proved to be high in the L2 group. Therefore, it would be worth investigating further its applicability to L1 productions. As we have seen, the low values of inter-rater agreement and reliability found in our data are possibly a consequence of the small range of scores provided by the raters. In order to overcome this limitation of the present study and to gain a more reliable picture of the scale's capability to measure native speakers' FA in writing, further investigations should take into account texts that are likely to be distributed along the whole scale, so that scores do not cluster at one end or the other. These should include texts written by informants having varying levels of educational attainment, which might allow raters to use all the levels of the scale.

5. Concluding remarks and pedagogical implications

In the light of the findings just discussed, we believe that the applicability of the FA scale proposed by Kuiken and Vedder (2017) to the productions of native speakers is worth exploring further, and within new contexts. More specifically, it would be interesting to see how this tool works when it is applied to pieces of academic writing. From a pedagogical point of view, the development of rating instruments that can measure and describe functional aspects of language proficiency in both L1 and L2 appears to be particularly desirable. Nowadays, most school and university teachers teach non-native speakers with different levels of proficiency in the L2 and native speakers with poor writing skills in their L1 simultaneously. It is well known, indeed, that in the last decades universities in several countries have been facing the challenge of ensuring the progression of a growing population of students with poor academic literacy skills in their L1s. Most of their weaknesses lay in the functional dimension of language use, like content selection and organization, or the application of genre features and register variation.

In the context of Italian universities, teachers have expressed concern about students' low levels of academic language proficiency since the 1990s (see, for instance, Lavinio & Sobrero, 1991). In the relevant literature, problems on functional aspects of language use are reported with both typical academic texts, such as essays and summaries, and communicative tasks that students need to accomplish in their university life, like writing emails to professors, motivation letters for exchange programs, or applications for internships. For example, in a corpus of essays produced by students of Italian literature at the University of Milan, Prada (2009) identified widespread weaknesses in the logical organization of arguments and in the use of genre conventions. More generally, the data revealed severe difficulties in producing comprehensible and communicatively effective texts. Andorno (2015) collected and analyzed 162 emails written to a professor by students of humanities in two universities of Northern Italy. Although the students seemed aware of the need to adopt a formal register, in some cases they were unable to use properly the linguistic devices associated to that register. In this context, the construct of FA seems to be of particular interest, as it might help better understand an area of research and intervention that requires careful attention with regard to university students' needs.

Another line of research with possible pedagogical import would be that of looking at the potential of the FA scale as a teaching and training tool. In the panel discussion with the raters conducted for this study, it emerged that using the scale to rate their peers' texts had helped them reflect on their own writing skills. Therefore, it would be interesting to explore the pedagogical use of the scale in writing courses, where the students might be guided by the descriptors in building and/or revising their texts.

Acknowledgments

The authors wish to thank all student raters for their evaluation of the data. We particularly thank Diego Cortés Velásquez and Federica Del Bono (Roma Tre University) for their assistance in data collection and organization. We also thank Sofía Martín Laguna (Universitat Jaume I), who developed the English version of the tasks used in this study.

References

Andorno, Cecilia (2014). Una semplice informalità? Le e-mail di studenti a docenti universitari come apprendistato di registri formali. In Massimo Cerruti, Elisa Corino, & Cristina Onesti (Eds.), *Lingue in contesto. Studi di linguistica e glottodidattica sulla variazione diafasica* (pp. 13–32). Edizioni Dell'Orso.

Bridgeman, Brent, Powers, Donald, Stone, Elizabeth, & Mollaun, Pamela (2012). TOEFL iBT speaking test scores as indicators of oral communicative language proficiency. *Language Testing, 29*(1), 91–108.

Cortés Velásquez, Diego, & Nuzzo, Elena (2017, April 19-21). *Assessing L1 functional adequacy: Can we use the same scale as for L2?* [Paper presentation]. Task-Based Language Teaching (TLBT) Conference 2017, Barcelona, Spain.

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

DeJong, Nivja H., Steinel, Margarita P., Florijn, Arjen F., Schoonen, Rob, & Hulstijn, Jan H. (2012a). The effect of task complexity on functional adequacy, fluency and lexical diversity in speaking performances of native and non-native speakers. In Alex Housen, Folkert Kuiken, & Ineke Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 121–142). John Benjamins.

DeJong, Nivja H., Steinel, Margarita P., Florijn, Arjen F., Schoonen, Rob, & Hulstijn, Jan H. (2012b). Facets of speaking proficiency. *Studies in Second Language Acquisition, 34*(1), 5–34.

Del Bono, Federica (2017). *Aspetti pragmatici nella valutazione di testi scritti: uno studio dell'adeguatezza funzionale in italiano L2*. Unpublished MA dissertation, Roma Tre University.

Doros, Gheorghe, & Lew, Robert (2010). Design based on intraclass correlation coefficients. *American Journal of Biostatistics, 1*(1), 1–8.

Faone, Serena, Pagliara, Francesca, & Vitale, Giuseppina (2017, April 19-21). *How to assess L2 information-gap tasks through functional adequacy rating scales* [Paper presentation]. Task-Based Language Teaching (TLBT) Conference 2017, Barcelona, Spain.

Fan, Xitao, & Sun, Shaojing (2014). Generalizability theory as a unifying framework of measurement reliability in adolescent research. *Journal of Early Adolescence*, *34* (1), 38–65.

Graham, Matthew, Milanowsky, Anthony, & Miller, Jackson (2012). *Measuring and promoting inter-rater agreement of teacher and principal performance ratings*. Center for Educator Compensation Reform (CECR).

Grice, Herbert P. (1975). Logic and conversation. In Peter Cole & Jerry L. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 41–58). Academic Press.

Gwet, Kilem Li (2014). *Handbook of inter-rater reliability* (4th ed.). Advanced Analytics, LLC.

Housen, Alex, Kuiken, Folkert, & Vedder, Ineke (Eds.). (2012). *Dimensions of L2 performance and proficiency. Complexity, accuracy and fluency in SLA*. John Benjamins.

James, Lawrence R., Demaree, Robert G., & Gerrit, Wolf (1984). Estimating within-group inter-rater reliability with and without response bias. *Journal of Applied Psychology,* 69, 85–98.

Kuiken, Folkert, Vedder, Ineke, & Gilabert, Roger (2010). Communicative adequacy and linguistic complexity in L2 writing. In Inge Bartning, Maisa Martin, & Ineke Vedder (Eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research* (pp. 81–100). Eurosla Monographs Series 1.

Kuiken, Folkert, & Vedder, Ineke (2017). Functional adequacy in L2 writing. Towards a new rating scale. *Language Testing, 34*(3), 321–336.

Kuiken, Folkert, & Vedder, Ineke (2018). Assessing functional adequacy of L2 performance in a task-based approach. In Naoko Taguchi & YouJin Kim (Eds.), *Task-based approaches to teaching and assessing pragmatics* (pp. 265–285). John Benjamins.

Lavinio, Cristina, & Sobrero, Alberto (1991). *La lingua degli studenti universitari*. La Nuova Italia.

McGraw, Kenneth O., & Wong, Seok P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30–46.

McNamara, Tim, & Roever, Carsten (2007). *Testing: The social dimension*. Blackwell Publishing.

Orrù, Paolo (2019). Misurare l'adeguatezza funzionale in testi scritti di apprendenti di italiano L2. *Italiano LinguaDue, 11*(1), 45–58.

Pagliara, Francesca (2017, October 6-7). *Valutare l'adeguatezza funzionale in produzioni scritte di studenti Marco Polo* [Paper presentation]. Dieci anni di didattica dell'italiano a studenti cinesi. Risultati, esperimenti, proposte, Siena, Italy.

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

Prada, Massimo (2009). Le competenze di scrittura e le interazioni comunicative attraverso lo scritto: problemi e prospettive per una didattica della scrittura. *Italiano LinguaDue, 1*, 232–278.

Révész, Andrea, Ekiert, Mmonika, & Torgersen, Eivind N. (2016). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics, 37*(6), 828–848.

Sato, Takanori (2012). The contribution of test-takers' speech content to scores on an English oral proficiency test. *Language Testing*, *29*(2), 223–241.

Stemler, Steven E., & Tsai, Jessica (2008). Best practices in inter-rater reliability. Three common approaches. In Jason W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 29–49). Sage Publications.

Vedder, Ineke (2016). Il ruolo dell'adeguatezza funzionale nelle produzioni scritte in lingua seconda: proposta per una scala di valutazione. In Elisabetta Santoro & Ineke Vedder (Eds.), *Pragmatica e interculturalità in italiano lingua seconda* (pp. 79–92). Franco Cesati.

Appendix A

RATING SCALES FOR FUNCTIONAL ADEQUACY (FROM KUIKEN & VEDDER 2017: 335-336)

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

6 The number of ideas is *extremely adequate* and they are very consistent to each other.

5 The number of ideas is *very adequate* and they are very consistent to each other.

4 The number of ideas is *adequate* and they are sufficiently consistent.

3 The number of ideas is *somewhat adequate*, even though they are not very consistent.

2 The number of ideas is *scarcely adequate* and the ideas lack consistency.

1 The number of ideas is *not at all adequate* and insufficient and the ideas are unrelated to each other.

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

6 *All* the questions and the requirements of the task have been answered.

5 *Almost all* the questions and the requirements of the task have been answered.

4 *Most (more than half)* of the questions and the requirements of the task have been answered.

3 *Approximately half* of the questions and the requirements of the task have been answered.

2 *Some (less than half)* of the questions and the requirements of the task have been answered.

1 *None* of the questions and the requirements of the task have been answered.

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

6 The text is *very easily comprehensible* and highly readable. The ideas and the purpose are clearly stated.

5 The text is *easily comprehensible* and reads smoothly. Comprehensibility is not an issue.

4 The text is *comprehensible*. Only a few sentences are unclear but are understood, without too much effort, after a second reading.

3 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.

2 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.

1 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.

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

6 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.

5 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.

4 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.

3 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.

2 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.

1 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.

Appendix B

TASK INSTRUCTIONS

TASK 1

The university newspaper organizes a writing contest every year. The jury consists of lecturers and students from the Literature and Language Department. This year's topic is as follows: Write about an anecdote of something that happened during a study trip you have been on (high school or university, e.g. Erasmus, exchange programmes).
Your text must include the following aspects:
When did you go on the study trip?
Where did you go?
Who did you go with?
What happened?
When did that happen?
Did you learn anything from this experience?
You have 30 minutes to write your text. Minimum number of words: 150 (about 15 lines). The use of a dictionary is not allowed.

TASK 2

You have rented out your house through a booking website to a young couple, Maria and Max, who are visiting your town. Because you are not going to be there when they arrive, you decide to leave them a note with instructions related to the house.
Your note must include the following information:
Start by welcoming your guests.
Explain where they can find basic things they may need during their stay.
Explain briefly how to use certain electronic devices.
Explain how they should leave the house.
You have 30 minutes to write your text. Minimum number of words: 150 (about 15 lines). The use of a dictionary is not allowed.

TASK 3

You are going to spend next year in a foreign city as part of an international exchange programme. The international office from your host university has offered you three options for accommodation. They are interested in knowing your decision and in what you base it on. Read the information below and decide on which of the choices you prefer. Write an e-mail to the Director of International Students, justifying your choice.
Your e-mail must include the following information:
Which accommodation you selected.
The reasons why you would make that choice taking into account life in that city.
The reasons why you did not select the other two options.

	International student residence	**Shared house**	**Studio**
Price	900€	500€	800€
Distance from university and city centre	20 min. by tram or 30 min cycling	35 min. overall (cycling or tram)	50 min. by bus + 10 min. walk or 40 min. cycling
Kitchen	Shared	Shared	Private
Bathroom	Private	Shared	Private
Internet	Available, no additional charge	Not available	Available, additional charge
Washing facilities	Available, additional charge	Available outside the building, additional charge	Included, no additional charge

You have 30 minutes to write your text. Minimum number of words: 150 (about 15 lines). The use of a dictionary is not allowed.

Elena Nuzzo, Università Roma Tre
elena.nuzzo@uniroma3.it

EN **Elena Nuzzo** is an associate professor of modern language instruction in the Department of Foreign Languages, Literatures, and Cultures at Roma Tre University. Her work focuses on research, education, and teacher training in the field of applied linguistics, with a specific interest in Italian as a second language. Her main areas of research include the practical applications of Speech Act theory in the learning and teaching of second languages, intercultural pragmatics, and task-based language teaching.

ES **Elena Nuzzo** es profesora asociada de Didáctica de las lenguas modernas en el Departamento de Lenguas, Literaturas y Culturas Extranjeras de la Università Roma Tre. Realiza actividades de investigación, enseñanza y formación de docentes en el campo de la lingüística aplicada, con especial interés hacia el italiano como L2. Sus áreas de investigación principales incluyen las aplicaciones prácticas de la teoría de los actos lingüísticos en el aprendizaje y enseñanza de las segundas lenguas, la pragmática transcultural y la enseñanza basada en tareas.

IT **Elena Nuzzo** Elena Nuzzo è professoressa associata in didattica delle lingue moderne presso il Dipartimento di Lingue, Letterature e Culture Straniere dell'Università Roma Tre. Svolge attività di ricerca, di insegnamento e di formazione dei docenti nel campo della linguistica applicata, con un interesse specifico per l'italiano come lingua seconda. Le sue principali aree di ricerca includono le applicazioni pratiche della teoria degli atti linguistici nell'ambito dell'apprendimento e dell'insegnamento delle lingue seconde, la pragmatica transculturale e la didattica basata sul compito.

Giuseppe Bove, Università Roma Tre
giuseppe.bove@uniroma3.it

EN **Giuseppe Bove** Giuseppe Bove is a professor of statistics in the Department of Education at Roma Tre University. He has also taught at the University of Siena and Sapienza University of Rome. He conducts research in the field of multivariate statistics, with a particular focus on data analysis methods (factor analysis and multidimensional scaling) and their application to the assessment of learning in national and international educational surveys.

ES **Giuseppe Bove** es profesor titular de estadística en el Departamento de Ciencias de la Educación de la Universidad Roma Tre. También enseño en la Universidad de Siena y en la Universidad de Roma "La Sapienza". Realiza actividades de investigación en el ámbito de la estadística multivariante, con especial referencia al método de análisis de datos (técnicas factoriales y de escala multidimensional) y sus aplicaciones a la evaluación del aprendizaje en encuestas educativas nacionales e internacionales.

IT **Giuseppe Bove** è professore ordinario di statistica nel Dipartimento di Scienze della Formazione dell'Università Roma Tre. Ha insegnato inoltre all'Università degli Studi di Siena e alla Sapienza Università di Roma. Svolge attività di ricerca nell'ambito della statistica multivariata, con particolare riferimento ai metodi dell'analisi dei dati (tecniche fattoriali e di *scaling* multidimensionale) ed alle loro applicazioni alla valutazione dell'apprendimento in indagini educative nazionali ed internazionali.