UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Measurement of functional adequacy in different learning contexts

Rationale, key issues, and future perspectives

Kuiken, F.; Vedder, I.

DOI
10.1075/task.00013.kui

Publication date
2022

Document Version
Final published version

Published in
TASK : Journal on Task-Based Language Teaching and Learning

License
CC BY

Link to publication

Citation for published version (APA):
Kuiken, F., & Vedder, I. (2022). Measurement of functional adequacy in different learning contexts: Rationale, key issues, and future perspectives. *TASK : Journal on Task-Based Language Teaching and Learning*, *2*(1), 8-32. https://doi.org/10.1075/task.00013.kui

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

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

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

Download date:09 Mar 2023

Measurement of functional adequacy in different learning contexts

Rationale, key issues, and future perspectives

Folkert Kuiken and Ineke Vedder
University of Amsterdam

Linguistic performance elicited by language tasks has generally been operationalized in terms of complexity, accuracy, and fluency (CAF). However, this study argues that assessment of L2 proficiency is impossible without taking into account the adequacy and efficacy of L2 performance. To that end, we developed a rating scale for measuring functional adequacy (FA). In order to investigate the validity, reliability, and applicability of the rating scale, a number of studies are reviewed in which FA was assessed by both expert and non-expert raters, in different learning contexts, for L2 and L1, involving various source and target languages, proficiency levels, task types and modalities. We discuss perspectives and challenges for the use of the FA rating scale, particularly with regard to task-based language assessment (TBLA).

Keywords: functional adequacy (FA), rating scale, reliability, validity, applicability, task-based language assessment (TBLA)

When measuring language performance, previous studies (Housen et al., 2012) have typically evaluated dimensions of complexity, accuracy, and fluency (CAF), whereas less attention has been devoted to the efficacy and appropriacy of language proficiency and learners' pragmatic abilities in a second language (L2). The importance of also assessing the communicative dimension as an essential component of L2 performance, in addition to CAF, has been emphasized by several authors, including De Jong et al. (2012a), Kuiken and Vedder (2014, 2017, 2018), Pallotti (2009) and Révész et al. (2016).

This paper argues that linguistic performance should not only be assessed by measures along the CAF-triad but also in terms of its functional adequacy (FA). From the perspective of task-based language teaching (TBLT) and task-based language assessment (TBLA), we consider FA to be a multi-layered con-

https://doi.org/10.1075/task.00013.kui | Published online: 20 June 2022
TASK 2:1 (2022), pp. 8–32. ISSN 2666-1748 | E-ISSN 2666-1756
Available under the CC BY 4.0 license. © John Benjamins Publishing Company

struct. In order to assess FA of L2 performance, we developed a rating scale (Kuiken & Vedder, 2017, 2018), which distinguishes four dimensions of the construct: Task Requirements, Content, Comprehensibility, and Coherence & Cohesion. Recently, several experimental studies have been published in which the FA rating scale has been employed for assessing the performance of various types of learners with different proficiency levels, who have been submitted to various task types and modalities. The goal of the study presented here is three-fold: (i) to explore the applicability of the FA rating scale in different learning contexts; (ii) to discuss the connection between FA and related issues (CAF, task type, language proficiency); (iii) to address future perspectives and challenges.

In what follows, we start by defining the construct of FA and the necessity to assess it, in addition to CAF, from a task-based perspective. We give an account of the theoretical underpinnings of the FA rating scale from the framework of TBLA, and we describe the dimensions of the FA rating scale. On the basis of a number of experimental studies that were conducted to test the rating scale, we then examine the applicability of the scale in relation to different source and target languages, task types, task modalities, and for different levels of L2 proficiency. Next, we explore the relationship between FA and CAF, task type, and proficiency level. In the concluding section of the paper, we discuss pedagogical issues for classroom practice: the use of the FA rating scale as a diagnostic tool for teachers and as an instrument for self-assessment and peer feedback by learners. The paper further addresses future perspectives and challenges for SLA research, including the effect of task modality on FA and the question of whether the FA rating scale can be used for interactional tasks.

Complexity, accuracy and fluency vs. functional adequacy

Linguistic performance elicited by language tasks has been generally assessed in terms of CAF. An array of measures has been proposed to assess these three dimensions (e.g., Bulté & Housen, 2012; Housen et al., 2012; Norris & Ortega, 2009; Wolfe-Quintero et al., 1998). Although there seems to be a growing consensus among researchers over which measures are best suited to assess linguistic performance, additional measures are still being proposed, for example, indices for morphological complexity (Pallotti & Brezina, 2019), phraseological complexity (Paquot, 2018, 2019), and propositional complexity (Vasylets et al., 2019).

As has been argued in various studies (De Jong et al., 2012a; Kuiken & Vedder, 2014, 2017, 2018; Pallotti, 2009; Révész et al., 2016), it is crucial to consider the functional dimension of oral and written L2 performance in addition to the linguistic dimension. As Ortega (2003, p. 494) states, "progress in all learner's lan-

guage ability for use may include syntactic complexification, but it also entails the development of discourse and sociolinguistic repertoires that the language user can adapt appropriately to particular communication demands".

The main claim underlying our research on FA is the conviction that the assessment of linguistic performance in L2 (and L1) is not possible without taking into consideration the efficacy and appropriacy of learners' performances (Kuiken & Vedder, 2014, 2017, 2018). If the primary goal of most language learning is to communicate successfully, L2 performance needs to be evaluated with both CAF indices as well as measures of FA in order to capture a wider array of learning outcomes associated with the accomplishment of real-world tasks. As TBLT is primarily concerned with language use in social contexts, the assessment of FA, viewed from a task-based perspective, is a key concern of TBLT and TBLA (Kuiken & Vedder, 2017, 2018).

A second theoretical underpinning of our research is the necessity to assess FA as a separate dimension from CAF (Pallotti, 2009). Thus far, few studies have investigated the relationship between CAF and FA in L2 performance, or between the growth of both dimensions (however, see Herraiz Martínez, 2018; Herraiz Martínez & Alcón Soler, 2019; Nuzzo & Bove, 2020; Révész et al., 2016; Strobl & Baten, this issue). Up until now, little is known about the specific linguistic features that contribute to the development of functionally adequate and appropriate speech, or about the relationship between CAF and FA for oral output.

Furthermore, there may be asymmetries in proficiency between the two components of learner performance, as illustrated by Example (1) below (Martín Laguna, forthcoming). In a decision-making task assigned to intermediate L2 university students of English, whose native language was Spanish/Catalan, participants were asked to write an email to the director of the Erasmus Exchange Office, indicating their first choice of accommodation for a study abroad year in a European country. Learners could choose from three available options (i.e., a shared house, the international students' residence or a studio), varying in price and facilities (e.g., free Wi-Fi, use of a kitchen, a private bathroom, a washing machine, proximity to the city center). They had 30 minutes to write the text; the use of a dictionary was not allowed.

(1) I am writing this letter in order to inform you about my decision regarding the three possible options of accommodation. I am a University student and I do not have a lot of money. My original idea was renting the shared house, since it is the cheapest one and, as I have already mentioned, I do not have too much money. However, there is not internet available there, and this is a problem, specially being abroad, as I could not text my family and friends. Another reason why I have not chosen the shared house is because the washing facilities are not included. Regarding the international student residence, it seems to be

> the most comfortable accommodation, being just 20 minutes far from the university and having everything included. However, it is too expensive, and I cannot afford such price. Yours faithfully, X

The letter written by the student is generally accurate and does not contain many errors. Syntactically, the text shows some variation, including a number of subordinate clauses introduced by conjunctions and different verb forms and tenses, while vocabulary choices are appropriate. However, with regard to the specific requirements of the task, the letter is less adequate. Rather than explicitly mentioning the preferred type of accommodation, the writer provides a couple of reasons for excluding two options, the shared house and the students' residence (no internet, no washing facilities, too expensive). The preferred choice (the studio) has to be inferred by the reader. There is no opening salutation of the addressee and the closure ("yours faithfully") is rather abrupt. Contrary to the linguistic dimension, the text is functionally inadequate and the argumentation is poor. The example thus shows that to acquire a complete picture of a learner's L2 proficiency, both CAF indices and measures for FA should be employed.

Task-based language assessment

FA as a task-related construct is considered in our research within the framework of TBLA. Norris (2016, p. 232) conceptualizes TBLA as "the elicitation and evaluation of language use (across all modalities) for expressing and interpreting meaning, within a well-defined communicative context (and audience), for a clear purpose, toward a valued goal or outcome". As emphasized by Norris (2016, p. 239), tasks have come to play an increasingly prominent role at many levels of and uses for language assessment.

TBLA highlights the task as a vehicle for eliciting authentic, goal-directed, and meaning-focused L2 performance (Long, 2015, 2016). Assessment tasks, like the writing of an email to the director of the Erasmus Exchange Office in Example (1), require L2 learners to perform form-function mapping processes, since communicative goals and speech acts should be verbalized in relation to a number of contextual features (e.g., the relationship with the addressee, the degree of imposition, communication modality, text type, and genre). Hence, given the goal-directedness of task performance, assessment tasks need to be designed carefully, in order to provide learners with the opportunity to engage in situated communicative interaction in which language is used (see González-Lloret, this issue).

The dilemma TBLA is confronted with is that on the one hand assessment tries to be as specific as possible and related to the particular target task learners

have been assigned; however, on the other hand – from a viewpoint of efficiency – more generic assessment procedures are preferred, which can be used for a whole array of language tasks. As noted by González-Lloret (2016, this issue), evaluation criteria should always be derived from the assessment task and the type of language. From a task-based perspective, these criteria should also take into account the goal that needs to be reached by performing the task and the interlocutors involved in fulfilling the task. A drawback of this approach is that if evaluative criteria are relative to a particular task, in reference to a specific real-world context, inferences may not be made beyond the specific target task and test context (Bachman, 2002). The challenge for TBLA is thus to find the right balance between these two contradictory forces. A potential avenue for exploring the application of TBLA for large-scale, high-stakes assessments may be the development in TBLA of a set of so called "prototypical", standardized tasks. Examples are: describing an object, person or picture; making a decision, solving a (communication) problem, etc. As has also been suggested by Norris (2016), assessment of such prototypical tasks has proven to offer a meaningful basis for eliciting and generalizing about learners' functional language proficiency, rather than mere knowledge of vocabulary and grammar rules. The launch of the TBLT Language Learning Task Bank (Gurzynski-Weiss & IATBLT, n.d.) is an important step in that direction. With the FA rating scale, we aimed to create a standardized and validated tool that can be employed for various (prototypical) task types.

Assessment of functional adequacy

In order to construct an instrument that can assess FA in a reliable and valid way, we followed the guidelines for developing a measurement framework of Norris and Ortega (2003), which consists of the following six steps: construct definition, behavior identification, task specification, behavior elicitation, observation scoring, data analysis (see also Révész & Brunfaut, 2021).

FA has been defined in various ways, such as successful information transfer (Upshur & Turner, 1995), pragmatic appropriateness (McNamara & Roever, 2007), text coherence and cohesion (Knoch, 2009), discursive practice and adequacy in oral communication (Ekiert et al., 2018; Révész et al., 2016), and successful task performance (De Jong et al., 2012a; b). Considered within the framework of TBLT and TBLA, in line with De Jong et al. (2012a; b) and inspired by the conversational maxims of Grice (1975), we define FA as a task-related construct in terms of successful task completion. The main focus in our definition is on the adequacy of L2 production in relation to a specific social context and target task, interlocutor, speech act, register, and task modality: a phone call to the dentist, a

reservation in a restaurant, a short note to a friend, or, in an academic context, an email to a thesis supervisor, a pitch presentation of a research project, etc. In terms of the Gricean maxims, the felicity and adequacy of the oral or written message transmitted by the speaker/writer is judged by the receiver in terms of the quantity, relation, manner, and quality of the information in the text (Kuiken & Vedder, 2017, 2018). We used the term *communicative* adequacy in an earlier study (Kuiken et al., 2010), following Pallotti's suggestions (2009). However, in more recent studies, we have adopted the more appropriate term *functional* adequacy given the task-related and goal-directed nature of the construct (Kuiken & Vedder, 2017, 2018).

The requirements of the FA rating scale are as follows: (i) deconstruction of relevant components of the construct; (ii) independence of descriptors of FA from linguistic descriptors in terms of CAF; (iii) objective and countable scale descriptors; (iv) applicability in various learning contexts (different types of learners, task types and modalities, expert and non-expert raters; (v) the possibility to use the scale for different source and target languages.

Designed as a six-point Likert scale, the FA rating scale (see Appendix A) was developed according to recommendations for Likert-type scale construction that have been summarized by Phakiti (2020). The scale comprises four dimensions: Task Requirements, Content, Comprehensibility, Coherence & Cohesion.

Task requirements

Have the requirements of the task been fulfilled successfully (e.g., genre, task type, speech acts, register, addressee)? This dimension focuses on the extent to which the task is completed in accordance with the particular genre, task type, speech acts, and register required in the message transmitted by the speaker/writer to the listener/reader, and the specific instructions and requirements of the task. In the case of Example (1), for instance, students were asked to write an email to an international student officer, indicating their choice of accommodation out of three options.

Content

Is the number of ideas provided in the text adequate and are they consistent to each other? Does the speaker/writer give as much information as is needed (are all crucial content elements mentioned?) in relation to the goals of the task (e.g., ordering food in a restaurant) and no more? This dimension takes into account the adequacy of the number of information units or concepts expressed in the text

and their possible thematic elaboration in terms of main and secondary content elements.

Comprehensibility

How much effort is required to understand text purpose and ideas? Following Bridgeman et al. (2012), this dimension takes into account the extent to which the message in the text is comprehensible for the intended listener/reader. Is the text immediately comprehensible or does the addressee need to reread or relisten to (certain fragments of) the text in order to understand what is meant?

Coherence & Cohesion

Is the text coherent and cohesive (e.g., use of strategies for coherence, cohesive devices)? This dimension focuses on the adequacy of the message of the speaker/writer in terms of the occurrence of coherent relationships (e.g., discourse markers, coherence breaks, number of repetitions; cf. Knoch, 2007, 2009, 2011) and cohesive ties (e.g., presence or absence of deictic elements, use of cohesive and anaphoric devices and strategies).

Inspired by the CEFR (Council of Europe, 2001), scale descriptors have been formulated for each of the six levels for these four dimensions. As has been emphasized in various studies (e.g., Becker, 2018), scale designers must make principled and justified decisions about the criteria to be included in the rating scale. It is essential that these criteria describe the relevant features that characterize the different levels of the constructs to be assessed, and that raters interpret them correctly and are able to apply them in a consistent manner. In designing the descriptors, we followed suggestions by Weigle (2002) and Luoma (2004) with respect to descriptors' consistency, length, clearness, concreteness, and explicitness (see also Kuiken & Vedder, 2021). For the complete scale, including the descriptors of the four dimensions for all six levels, see Appendix A.
In the following section, we describe four studies that were conducted to test the reliability and validity of the FA rating scale. Next, in order to assess the applicability of the scale, a number of studies are presented in which the FA rating scale has been employed in different contexts: for both oral and written tasks, various task types and modalities, different source and target languages, and different levels of L2 proficiency (A2-C1). Based on the outcomes of these studies, we discuss the following questions:

1. To what extent can the FA rating scale be applied in different learning contexts?
2. What is the relationship between FA and CAF, task type and proficiency level?

Testing the FA rating scale

In order to test the reliability and validity of the FA rating scale, four studies were conducted. The first two (Kuiken & Vedder, 2014; Kuiken et al., 2010) are pilot studies in which a preliminary version of the FA rating scale was used. As a result of the outcomes and interviews with the raters involved in these studies, a modified version was presented in Kuiken and Vedder (2017), based on the analysis of written data; subsequently, the scale was tested for oral data (Kuiken & Vedder, 2018). For an overview of these studies, with respect to participant characteristics, task type, raters, and modalities, see Appendix B.

In Kuiken et al. (2010), a preliminary, global version of the FA rating scale consisting of six levels was piloted on students of Dutch L2 ($N=34$), Italian L2 ($N=42$), and Spanish L2 ($N=27$), who performed two written decision-making tasks. 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 on the front page of their favorite newspaper. All data were collected at a university in the Netherlands. The Dutch L2 participants came from various language backgrounds while the Italian L2 and Spanish L2 students were native speakers of Dutch. The proficiency level of the three language groups varied from A2 to B1 according to the CEFR (Council of Europe, 2001). The tasks were judged by expert raters, who were all native speakers of the target languages (Dutch: $N=4$; Italian: $N=3$; Spanish: $N=3$). Interrater reliability scores for FA, as measured by Cronbach's α, were acceptable, ranging from 0.70 to 0.78.

The preliminary version of the FA rating scale was used again in a similar study (Kuiken & Vedder, 2014), based on the same data from the learners of Dutch ($N=32$) and Italian ($N=39$).[1] Data from native speakers of Dutch ($N=17$) and Italian ($N=18$) who performed the same tasks were used as a comparison group. As in Kuiken et al. (2010), expert raters (experienced L2 teachers) were asked to judge the performance of the learners (Dutch: $N=4$, Italian: $N=3$). After a short

1. Participant numbers in Kuiken and Vedder (2014) differ slightly from those in Kuiken et al. (2010), because in the former study, participants who had not performed both tasks were excluded.

training session, these interrater reliability scores, as measured by Cronbach's α, varied from 0.72 to 0.79 for the L2 learners and from 0.70 to 0.90 for the L1 learners. In view of the validation of the rating scale, so-called cognitive interviews (Phakiti, 2020) were held with the raters, in order to find out how they interpreted the scale and whether the descriptors were clear. The interviews took place in the form of a three-hour panel discussion, one with the raters of Dutch and one with those of Italian. During these sessions, raters of both languages reported that their fundamental concern was whether writers managed to get their message across. What appeared most important to them were text comprehensibility and rhetorical organization.

Based on these two studies, the preliminary FA rating scale was adapted to its present form and split into the four dimensions presented in the preceding section (Kuiken & Vedder, 2017, 2018). Because one of the requirements for the rating scale was that it could be used by both expert and non-expert raters, the written data discussed above (Kuiken & Vedder, 2014; Kuiken et al., 2010) were now presented to non-expert raters: all of them were university students of about the same age as the participants involved in the study (Dutch: $N=4$, Italian: $N=4$). Interrater agreement in terms of intraclass correlation coefficients varied from acceptable (0.73) to excellent (0.94).

In order to investigate the applicability of the new FA rating scale when assessing oral data, the same decision-making tasks were presented in an oral task modality to learners of Dutch L2 ($N=22$, level A2-B2) and Italian L2 ($N=26$, level A2-C1). In this study (Kuiken & Vedder, 2018), non-expert raters were asked to assess FA by means of the FA rating scale, which had been slightly adapted for the assessment of the oral tasks: designations such as "text", "writer" and "reader" used in the scale for writing were replaced by "performance", "speaker" and "listener" in the scale for speaking. All raters were advanced university students and native speakers of the target language (Dutch: $N=4$; Italian: $N=4$). Intraclass correlations among the raters on the oral performance varied for the two languages from good (0.86) for Dutch to excellent (0.93) for Italian.

A couple of issues to be kept in mind emerged from these studies. Firstly, the necessity of rater training prior to the use of the FA rating scale. This point has also been emphasized in several other studies, both with respect to rating scales in general (Pill & Smart, 2020) and in particular with respect to the assessment of pragmatic features (González-Lloret, this issue) and FA (Faone & Pagliara, 2017; Kuiken & Vedder, 2014). In a study on the communicative competence of German learners of English (Timpe, 2013; Timpe-Laughlin, 2018), it was found that while lexical or grammatical mistakes were easily perceived by raters as L2 deficiencies, they appeared to struggle with rating the appropriacy of L2 performance. In order to familiarize raters with scale dimensions, levels and descriptors, rater training is

thus crucial, possibly even more so in the case of the assessment of FA, as it will lead to higher validity and reliability (Rezaei & Lovorn, 2010).

A related issue which challenged raters was the use of the FA rating scale for both native speakers and L2 learners, in one and the same study. When raters were asked to judge both groups, we found that there was less unanimity between raters in L1 compared to L2 (Kuiken & Vedder, 2014). This result is in line with Schoonen (2005), who concluded that raters are more consistent in judging L2 performance than L1 output. This may be due to the fact that the FA scale was inspired by the CEFR descriptors, which were designed for second language learners and not for native speakers.

Another outcome of the debriefing sessions was that raters appeared to consider the Coherence & Cohesion dimension the most difficult to assess. As its name suggests, this scale dimension combines two different aspects. Although coherence and cohesion are often connected in the assessment and teaching literature (which is the reason we merged them), the two notions are theoretically distinct; it is possible that a text is coherent without the use of any connectives or anaphoric devices. Furthermore, coherence and cohesion may be especially hard to assess in speaking, where direct referral to what has been stated before is harder than in writing (Kuiken & Vedder, 2012).

As such, these first studies in which the FA rating scale was tested indicate that the scale seems to be a reliable and valid tool, taking into consideration the participants involved in the four studies (L2 and L1 Dutch, Italian and Spanish), task type (decision-making), task modality (speaking and writing), and raters (expert and non-expert). It should, however, be noted that the scale was used exclusively for adult, highly educated L2 learners of Dutch, Italian and Spanish, who were subjected to one type of task. In the following section, we will consider the applicability of the scale in other studies conducted in different learning contexts.

Studies in which the FA rating scale has been used

Since the launch of the FA rating scale, the instrument has been used by several researchers in various settings, particularly for learners of English L2 and Italian L2 with different first languages (for an overview, see Appendix C). For English L2, FA has been assessed in the American context, involving 80 Japanese and Spanish English L2 learners (Révész et al., 2016; Ekiert et al., 2018, this issue). The speaking performance of the participants was assessed by both expert and non-expert raters. Learners were equally divided over four proficiency groups (low-intermediate, intermediate, low-advanced and advanced, with 20 learners in each group), while a native speaker group ($N = 20$) served as a comparison group. The

learners were subjected to various task types: a complaint about a catering service, refusing a suggestion by a teacher, telling a story based on pictures, giving advice based on a radio commentary, and summarizing information from a lecture.

Additionally, the FA rating scale has been used in Spain in three studies for assessing the written performance of Spanish/Catalan learners of English L2. Martín Laguna (forthcoming) gave beginner ($N=25$) and advanced ($N=25$) learners three different writing tasks. In the narrative task, the participants had to describe an episode that occurred during a study trip abroad. The instruction task consisted of writing instructions to a couple about an apartment that the students would rent for a week. In the decision-making task, an email had to be written to an international student officer about the choice of residence type, out of three options, during a study abroad program (see Example (1)). Non-expert raters were asked to assess the FA of the resulting texts. Herraiz Martínez (2018) and Herraiz Martínez and Alcón Soler (2019) asked expert raters to assess the FA of motivation letters to participate in Erasmus exchange programs or to conduct internships, written by 102 learners of English L2 with varying proficiency levels (A2-C1).

The applicability of the FA rating scale has also been examined for Italian L2 in a number of studies in which (mostly) non-expert raters assessed the written performance of L2 learners of Italian with diverse language backgrounds (including Chinese, Dutch and Hungarian) and various proficiency levels (A2-C1). In most of these studies (Del Bono, 2019, 2020; De Meo et al., 2019; Faone & Pagliara, 2017; Orrù, 2019; Orrù & Foti, 2020), the same three tasks were used that had already been employed in the study of Martín Laguna (forthcoming): decision-making (choosing an accommodation during an exchange programme), narration (describing an episode during a study trip abroad) and instruction (briefing a couple who will look after a house). In contrast to these studies, which all involved adult L2 learners, Pallotti (2017a, b, c, this issue) used the FA rating scale for assessing the narration skill of 217 primary school children (grades 3–5), of which 153 were monolingual L1 speakers of Italian, while 64 were multilinguals, having Italian as a second or additional language. All children undertook a speaking task in which they were asked to recount a short episode of Charlie Chaplin's silent film *Modern Times*, requiring them to move from the description of background states to the narration of events. Nuzzo and Bove (2020), who also employed the three tasks developed by Martín Laguna, investigated the pedagogical use of the FA rating scale in both an L2 ($N=20$) and an L1 ($N=20$) writing context. Nuzzo and Bove (this issue) also explored the applicability of the FA rating scale as a teaching tool for L1 writing instruction, based on data collected from 30 Italian university students at MA level, who were asked to write a motivation letter to apply as trainers for in-service secondary school teachers.

Finally, Strobl and Baten (this issue) investigated the relationship between FA and CAF in German L2 writing. They asked three expert raters to assess two narrative writing tasks (a personal narrative to a friend), related to the study abroad expectations and experiences of 30 Belgian Dutch-speaking learners of German L2, who participated in a three- to four-month study abroad program.

As demonstrated by these studies, the FA rating scale has been applied by both expert and non-expert raters for assessing the linguistic performance of diverse language learners (children and adults, L1 and L2 learners) in a variety of oral and written tasks, in several target and source languages, and with different levels of proficiency (A2-C1), next to native speakers). Nevertheless, it remains important to reconsider reliability and validity in these studies, as they may be affected by type and number of participants, task type, and time and place at which data collection takes place.

Outcomes

Based on these studies in which the FA rating scale has been used, we report on the main outcomes of the studies that have been presented and the way in which the FA rating scale was employed. We then explore the relationship between FA on the one hand and CAF, task type, and proficiency level on the other.

Use of the FA rating scale

Researchers who have made use of the FA rating scale have either employed it as it was originally developed or have made some minor adaptations to the instrument. Studies that have applied the original FA rating scale include Faone and Pagliara (2017), Del Bono (2019, 2020), Orrù (2019); Orrù and Foti (2020), Nuzzo and Bove (2020, this issue), and Martín Laguna (forthcoming). Strobl and Baten (this issue) reduced the four dimensions of the rating scale to three by combining the dimensions of Task Requirements and Content into one dimension, which they labelled Content & Topic Development. Others have extended the scale. While Coherence & Cohesion were combined into the same dimension in the FA rating scale, Herraiz Martínez (2018) and Herraiz Martínez and Alcón Soler (2019) separated Coherence from Cohesion, resulting in a scale with five dimensions. Pallotti (2017a, b, c, this issue) left out Task Requirements but added the CEFR scale for Coherence & Cohesion. A task-independent scale, supplemented by task-dependent content points, was used by Ekiert et al. (2018), Ekiert et al. (this issue). Révész et al. (2016). The scale consisted of seven levels, with descriptors related to whether the speaker addressed and supported the task-specific con-

tent points with sufficient detail, was easy or difficult to understand, delivered the message in a clear and effective manner, and took into account the communicative situation. Finally, some researchers calculated a composite FA score, based on the average of the four subscales.

Studies that have calculated interrater reliability scores concluded that they varied from acceptable to excellent, at least as far as L2 learners were concerned – cf. Cronbach's α scores for L2 learners: Del Bono (2019): 0.73–0.95, Faone and Pagliara (2017): 0.89–0.93, Orrù (2019), Orrù and Foti (2020): 0.78–0.88. More variation was found for interrater agreement – cf. intraclass correlation scores for L2 learners: Del Bono (2019): 0.30–0.79, Faone and Pagliara (2017): 0.40–0.56, Orrù (2019), Orrù and Foti (2020): 0.77–0.84. A possible explanation for this variation between studies might be that interrater agreement among expert and non-expert raters was not sufficiently high, as demonstrated, for instance, in the study by Faone and Pagliara (2017), where the relatively low interrater agreement was likely to be attributed to the fact that their raters had not received any training on how to use the FA rating scale. In line with our recommendations mentioned earlier, the authors emphasize that rater training is crucial for achieving acceptable interrater agreement scores (see also Pill & Smart, 2020; Rezaei & Lovorn, 2010). Lower interrater agreement scores were also found in the study by Nuzzo and Bove (2020), in which both L2 learners and native speakers were rated by means of the FA rating scale (intraclass correlations for L2: 0.24–0.63; for L1: 0.02–0.43). As also observed by Kuiken and Vedder (2014), native speakers, when combined with L2 learners in the same experiment, appeared to receive higher scale level scores. This range restriction of the rating instrument resulted in low interrater correlations and low alpha values. Nevertheless, this does not seem to hold when only native speakers are involved, as shown by Nuzzo and Bove (this issue). Again, it should be emphasized that whenever the scale is adapted or used for a different audience, it will need to be subjected to further validation.

Relationship with CAF, task type, and language proficiency

Several studies have looked at the relationship of FA with CAF, task type, and language proficiency. The impact of language proficiency on FA was investigated by Herraiz Martínez (2018) and Herraiz Martínez and Alcón Soler (2019). In a pretest-posttest-delayed posttest design, comprising one academic year, 102 Spanish/Catalan learners of English L2 were asked to write motivation letters in order to be admitted to university. Proficiency level was assessed by means of the Oxford Quick Placement Test, focusing on language use and listening. Scores were converted into CEFR levels. At the beginning of the academic year, learners at level B2 scored significantly better on all dimensions of FA than their peers at level

B1. However, these differences between the two groups were not maintained over time. Although all students tended to score higher throughout the year, Content was the only dimension in which the B2 group scored significantly better in the posttest, whereas significant differences were no longer found in the delayed posttest. These results suggest that on the posttest, there were no differences in FA between the two groups, regardless of the learners' initial proficiency levels.

Nuzzo and Bove (2020) submitted 20 learners of Italian L2 and 20 native speakers of Italian to Martín Laguna's three writing tasks which have been mentioned above (narrative, instruction, decision-making). They investigated how scores on FA correlated with general levels of proficiency as measured by a C-test. In the L2 group, global FA scores, operationalized as the composite score of the four subdimensions, correlated significantly with C-test scores, particularly for the instruction task (Pearson's product moment correlation 0.87). Moderately high correlations were observed for the narrative (0.70) and the decision-making (0.61) tasks. These results suggest that overall proficiency as measured by the C-test is not independent from FA, and that an association can be observed, although the correlation varies across task types.

Révész et al. (2016) investigated to what degree task type and proficiency level influence the extent to which CAF measures predict FA. The authors subjected 80 English L2 learners, divided over four equal groups of different proficiency levels, to five speaking tasks (complaint, refusal, narration, advice, and summary). One of the findings was that the subordination ratio of speakers appeared to function as a predictor of FA. This contradicts the results of Kuiken et al. (2010), who did not identify subordination complexity as a significant predictor of FA. This different outcome may be due to a difference in task modality, as participants in the latter study were given a writing task. In the same paper, Révész et al. concluded that repair fluency was the only CAF measure that showed differential impact on FA depending on proficiency. Higher scores on FA were found to be associated with lower incidence of false starts in the advanced L2 users' speech. Task type was not found to moderate the relationship between FA and the CAF measures. However, in a follow-up study based on the same participants (Ekiert et al., 2018), in which the researchers focused on the complaint, refusal, and advice tasks, their earlier result (no effect of task type) was confirmed for highly proficient learners, but not for less proficient learners, who appeared to struggle especially with the refusal task. Following up on this, Ekiert et al. (this issue) found, based on the data of the complaint and refusal tasks performed by a subset of the learners (ten participants from each of the four proficiency levels), that the fewer silent pauses L2 speakers produced between clauses, the more functionally adequate they were perceived.

A relationship between FA and fluency has also been found in written performance. In narrative writing, Strobl and Baten (this issue) observed strong correlations between the number of words written and lexical sophistication on the one hand and scores on Content and Coherence & Cohesion on the other. Comprehensibility was weakly associated with accuracy (computed by the number of error-free clauses per total number of clauses) and mean word length. No relationship between FA and syntactic complexity could be established.

The overall conclusion we can draw from these studies is that they have resulted in mixed findings regarding the connection between FA and CAF, FA and proficiency level, and FA and task type, and that these relationships need to be further investigated.

Discussion

The various studies in which the FA rating scale was employed have shown that its use is not limited to particular types of learners or tasks and that, in terms of applicability, its scope is sufficiently broad. The scale can be utilized – occasionally with some minor adaptations – for language learners of different ages, levels, source and target languages, and for different types of oral and written tasks. It should, however, be pointed out that the reliability and validaty of the scale should be reconsidered whenever it is employed and/or adapted in other contexts than the ones in which it has been used so far.

FA has to be viewed as a key construct for assessing language performance in general and task-based language performance in particular. Although dimensions of FA and CAF across proficiency levels appear to be connected to some degree, the overview presented in this paper shows that the two constructs are fundamentally distinct, which makes it necessary to evaluate them separately. Therefore, in language assessment, we should consider both the CAF and FA dimension, extending CAF to CAFFA (see also Pallotti, this issue; González-Lloret, this issue). Alongside studies on the development of FA in relation to CAF, task type and language proficiency (Strobl & Baten, this issue; Ekiert et al., 2018, this issue; Herraiz Martínez, 2018; Herraiz Martínez & Alcón Soler, 2019; Nuzzo & Bove, 2020; Révész et al., 2016), further research in this intriguing research area is essential.

Based on the outcomes of the studies that have made use of the FA rating scale, we may conclude that the scale is user-friendly, as both expert and non-expert raters were able to use it after one or two training sessions. The importance of rater training should nonetheless be emphasized, since rater training has been found to increase both the validity and reliability of the test instrument (see also

Becker, 2018; Pill & Smart, 2020). As shown in the study by Faone and Pagliara (2017), lack of training, conversely, may result in low interrater agreement.

As far as pedagogical implications are concerned, the user-friendliness of the FA rating scale leads to the question of whether, as well as how, the scale can also be used in classroom practice. The rating scale allows teachers to provide their students with more focused feedback on FA. Scores obtained by L2 learners regarding the functional dimension of their oral and written performances may more precisely indicate strengths and weaknesses of each learner, providing teachers useful information in terms of teaching targets and feedback. In this way, the FA rating scale may serve as a diagnostic instrument and give learners insight into their own abilities by showing them which adequacy levels they already have reached and which will be their next learning targets on which to focus. The scale may also be employed as a tool for self-assessment by learners and/or peer feedback (see Nuzzo & Bove, this issue). Further research is necessary in order to establish to what extent the FA rating scale can be applied for these pedagogical purposes.

This brings us to future perspectives and remaining challenges. The first area that requires further research is the effect of task modality on FA. As demonstrated in the previous sections, the FA rating scale has been used successfully to assess FA in both writing and speaking tasks. However, it is not yet fully clear to what extent task modality affects assessment of FA. It is easy to imagine that Comprehensibility may be evaluated differently in written texts than in oral speech, where raters may be influenced or distracted by pronunciation, intonation, rhythm, and pitch. Studies which have investigated the influence of task modality on CAF in L2 performance have produced mixed results (see e.g., Kuiken & Vedder, 2012). How far this also holds for the functional dimension of L2 performance remains to be seen. As a large number of tasks can be performed in both task modalities, research is encouraged which compares learners' FA in speaking tasks with their performance in the same (or similar) writing tasks, at various proficiency levels.

Another important issue that should be addressed in future research is the question whether the FA rating scale can also be used for interactional tasks (see also González-Lloret, this issue). So far, the FA scale has been employed exclusively for the assessment of monologic tasks. As nearly all communication takes place in an interactional setting, keeping in mind a recent study by Pallotti (2019) in which a tool for measuring interactional competence was proposed, one may wonder what this implies for the FA rating scale. It may be that new scale descriptors have to be added (e.g., adequacy of turn taking, topic switches, etc.) or that the scale has to be extended with one or more new dimensions.

In sum, assessing language performance in terms of both CAF and of FA, has proven to be an important step forward. Future research should focus on remaining challenges and new perspectives.

Funding

Open Access publication of this article was funded through a Transformative Agreement with University of Amsterdam.

References

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing*, *19*(4), 453–476. https://doi.org/10.1191/0265532202lt2400a

Becker, A. (2018). Not to scale? An argument-based inquiry into the validity of an L2 writing scale. *Assessing Writing*, *37*, 1–12. https://doi.org/10.1016/j.asw.2018.01.001

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. https://doi.org/10.1177/0265532211411078

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency. Complexity, accuracy and fluency in SLA* (pp. 21–46). John Benjamins. https://doi.org/10.1075/lllt.32.02bul

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. 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). John Benjamins. https://doi.org/10.1075/lllt.32.06jon

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. https://doi.org/10.1017/S0272263111000489

Del Bono, F. (2019). Aspetti pragmatici nella valutazione di testi scritti: Uno studio sull'adeguatezza funzionale in italiano L2. In: E. Nuzzo, & I. Vedder (Eds.), *Lingua in contesto. La prospettiva pragmatica. Studi AItLA 9* (pp. 231–244). Associazione Italiana di Linguistica Applicata (AitLA).

Del Bono, F. (2020). L'utilizzo delle scale dell'adeguatezza funzionale su testi narrativi in L2: Uno studio esplorativo sugli effetti del task design. In: E. Nuzzo, E. Santoro, & I. Vedder (Eds.), *Valutazione e misurazione delle produzioni orali e scritte in italiano lingua seconda* (pp. 71–82). Franco Cesati Editore.

De Meo, A., Maffia, M., & Vitale, G. (2019). La competenze scritta in italiano L2 di apprendenti vulnerabili. Due scale di valutazione a confronto. *EL.LE*, *8*(3), 637–654. https://doi.org/10.30687/ELLE/2280-6792/2019/03/007

Ekiert, M., Lampropoulou, S., Révész, A., & Torgersen, E. (2018). The effects of task type and L2 proficiency on discourse appropriacy in oral task performance. In N. Taguchi, & Y-J. Kim (Eds.), *Task-based approaches to assessing pragmatics* (pp. 247–264). John Benjamins. https://doi.org/10.1075/tblt.10.10eki

Faone, S., & Pagliara, F. (2017). *How to assess L2 information-gap tasks through FA rating scales*. Paper presented at TBLT 2017.

González-Lloret, M. (2016). *A practical guide to integrating technology into task-based language teaching*. Georgetown University Press. https://doi.org/10.6035/LanguageV.2017.9.9

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

Gurzynski-Weiss, L., & IATBLT (n.d.). *The TBLT Language Learning Task Bank*. https://tblt.indiana.edu

Herraiz Martínez, A. (2018). *Functional adequacy: The influence of English-medium instruction, English proficiency, and previous language learning experiences*. Doctoral dissertation, Universitat Jaume I, Castellón de la Plana.

Herraiz Martínez, A., & Alcón Soler, E. (2019). Pragmatic outcomes in the English-medium instruction context. *Applied Pragmatics*, *1*(1), 68–91. https://doi.org/10.1075/ap.00004.her

Housen, A., Kuiken, F., & Vedder, I. (Eds.). (2012). *Dimensions of L2 performance and proficiency. Complexity, accuracy and fluency in SLA*. John Benjamins. https://doi.org/10.1075/lllt.32

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. https://doi.org/10.1016/j.asw.2007.07.002

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, *26*(2), 275–304. https://doi.org/10.1177/0265532208101008

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. https://doi.org/10.1016/j.asw.2011.02.003

Kuiken, F., & Vedder, I. (2012). Speaking and writing tasks and their effects on second language performance. In S.M. Gass, & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 364–377). Routledge. https://doi.org/10.4324/9780203808184

Kuiken, F., & Vedder, I. (2014). Rating written performance: What do raters do and why? *Language Testing*, *31*(3), 329–348. https://doi.org/10.1177/0265532214526174

Kuiken, F., & Vedder, I. (2017). Functional adequacy in L2 writing. Towards a new rating scale. *Language Testing*, *34*(3), 321–336. https://doi.org/10.1177/0265532216663991

Kuiken, F., & Vedder, I. (2018). Assessing functional adequacy of L2 performance in a task-based approach. In N. Taguchi, & Y-J. Kim (Eds.), *Task-based approaches to assessing pragmatics* (pp. 265–286). John Benjamins. https://doi.org/10.1075/tblt.10.11kui

Kuiken, F., & Vedder, I. (2021). Scoring approaches: Scales/rubrics. In P. Winke, & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 125–134). Routledge.

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). European Second Language Association.

Long, M. H. (2015). *Second language acquisition and task-based language teaching*. Wiley Blackwell.

Long, M. H. (2016). In defense of tasks and TBLT: Nonissues and real issues. *Annual Review of Applied Linguistics*, *36*, 5–33. https://doi.org/10.1017/S0267190515000057

Luoma, S. (2004). *Assessing speaking*. Cambridge University Press. https://doi.org/10.1017/CBO9780511733017

Martín Laguna, S. (forthcoming). *Testing functional adequacy in L2 writing across languages, levels and tasks*. Universitat Jaume I, Castellón de la Plana.

McNamara, T., & Roever, C. (2007). *Testing: The social dimension*. Blackwell. https://doi.org/10.1111/j.1473-4192.2006.00117.x

Norris, J. M. (2016). Current uses for task-based language assessment. *Annual Review of Applied Linguistics*, *36*, 23–244. https://doi.org/10.1017/S0267190516000027

Norris, J. M., & Ortega, L. (2003). Defining and measuring SLA. In C. J. Doughty, & M. H. Long (Eds.), *The handbook of seond language acquisition* (pp. 717–761). Blackwell. https://doi.org/10.1002/9780470756492.ch21

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, *30*(4), 555–578. https://doi.org/10.1093/applin/amp044

Nuzzo, E., & Bove, G. (2020). Assessing functional adequacy across tasks: A comparison of learners' and speakers' written texts. *E-JournALL*, *7*(2), 9–27. https://doi.org/10.21283/2376905X.12.175

Orrù, P. (2019). Misurare l'adeguatezza funzionale in testi scritti di apprendenti di italiano L2. *Italiano LinguaDue*, *1*, 45–58. https://doi.org/10.13130/2037-3597/11843

Orrù, P., & Foti, E. (2020). *Coerenza e coesione nella valutazioni dell'adeguatezza funzionale: Un confronto tra i giudizi dei valuatori*. In: E. Nuzzo, E. Santoro, & I. Vedder (Eds.), *Valutazione e misurazione delle produzioni orali e scritte in italiano lingua seconda* (pp. 83–92). Franco Cesati Editore.

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. https://doi.org/10.1093/applin/24.4.492

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, *30*(4), 590–601. https://doi.org/10.1093/applin/amp045

Pallotti, G. (2017a). Applying the interlanguage approach to language teaching. *IRAL*, *55*(4), 393–412. https://doi.org/10.1515/iral-2017-0145

Pallotti, G. (2017b). Osservare l'interlingua. Percorsi di educazione linguistica efficace per ridurre le diseguaglianze. In M. Vedonelli (Ed.), *L'italiano dei nuovi italiani. Atti del XIX Convegno Nazionale GISCEL* (pp. 505–520). Aracne.

Pallotti, G. (2017c). Une application des recherches sur l'interlangue aux contextes d'enseignement. *Le Français dans le monde*, *61*, 109–120.

Pallotti, G. (2019). Assessing tasks: The case of interactional difficulty. *Applied Linguistics*, *40*(1), 176–197. https://doi.org/10.1093/applin/amx020

Pallotti, G., & Brezina, V. (2019). Morphological complexity in written L2 texts. *Second Language Research*, *35*(1), 99–119. https://doi.org/10.1177/0267658316643125

Paquot, M. (2018). Phraseological competence: A missing component in university entrance language tests? Insights from a study of EFL learners' use of statistical collocations. *Language Assessment Quarterly*, *15*(1), 29–43. https://doi.org/10.1080/15434303.2017.1405421

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research*, *35*(1), 121–145. https://doi.org/10.1177/0267658317694221

Phakiti, A. (2020). Likert-type scale construction. In P. Winke, & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 102–114). Routledge. https://doi.org/10.4324/9781351034784-12

Pill, J., & Smart, C. (2020). Rating: Behavior and training. In P. Winke, & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 135–144). Routledge. https://doi.org/10.4324/9781351034784-15

Révész, A., Ekiert, M., & Torgersen, E. (2016). The effects of complexity, accuracy and fluency on communicative adequacy in oral task performance. *Applied Linguistics*, *37*(6), 828–848. https://doi.org/10.1093/applin/amu069

Révész, A., & Brunfaut, T. (2021). Validating assessments for research purposes. In P. Winke, & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 21–32). Routledge.

Rezaei, A.R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing*, *15*(1), 18–39. https://doi.org/10.1016/j.asw.2010.01.003

Schoonen, R. (2005). Generalizability of writing scores. An application of structural equation modeling. *Language Testing*, *22*(1) 1–30. https://doi.org/10.1191/0265532205lt2950a

Timpe, V. (2013). *Assessing intercultural communicative competence. The dependence of receptive sociopragmatic competence and discourse competence on learning opportunities and input*. Peter Lang.

Timpe-Laughlin, V. (2018). Pragmatics in task-based language assessment. Opportunities and challenges. In N. Taguchi, & Y-J. Kim (Eds.), *Task-based approaches to assessing pragmatics* (pp. 288–304). John Benjamins. https://doi.org/10.1075/tblt.10.12tim

Upshur, J.A., & Turner, C.E. (1995). Constructing rating scales for second language tests. *ELT Journal*, *49* (1), 3–12. https://doi.org/10.1093/elt/49.1.3

Vasylets, O., Gilabert, R., & Manchón, R.M. (2019). Differential contribution of oral and written modes to lexical, syntactic and propositional complexity in L2 performance in instructed contexts. *Instructed Second Language Acquisition*, *3*(2), 206–227. https://doi.org/10.1558/isla.38289

Weigle, S. (2002). *Assessing writing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732997

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity*. University of Hawai'i Press.

Appendix A. FA rating scale for functional adequacy (for writing)

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

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

Content: Is the number of ideas provided in the text adequate and are they consistent to each other?

1	2	3	4	5	6
The number of ideas is not at all adequate and the ideas lack consistency.	The number of ideas is scarcely adequate and the ideas are hardly consistent.	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 with each other.	The number of ideas is extremely adequate and they appear very consistent with each other.

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. Purposes and ideas 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 purposes and ideas.	The text is somewhat comprehensible and 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. Purposes and ideas are clearly stated.	The text is easily comprehensible and reads smoothly; comprehensibility is not an issue. Purposes and ideas are clearly stated.	The text is very easily comprehensible and highly readable; the purposes and ideas are clearly stated.

Coherence & Cohesion: Is the text coherent and cohesive (e.g., use of strategies for coherence, cohesive devices)?

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 devices. 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 exceptional coherence by integrating new ideas into the text with connectives or connective phrases. Anaphoric devices are used regularly. There are few instances 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.

Appendix B. Studies for testing out the FA rating scale

Studies (in chronological order)	Language						Raters		Modality	
	NNS									
	L2	L1	Level	*N*	NS	Task type	Expert	Nonexpert	Writing	Speaking
Kuiken, Vedder & Gilabert (2010)	Dutch	Various	A2-B1	*N* = 34		Decision	*N* = 4		X	
	Italian	Dutch	A2-B1	*N* = 42			*N* = 3			
	Spanish	Dutch	A2-B1	*N* = 27			*N* = 3			
Kuiken & Vedder (2014)	Dutch	Various	A2-B1	*N* = 32	*N* = 17	Decision	*N* = 4		X	
	Italian	Dutch	A2-B1	*N* = 39	*N* = 18		*N* = 3			
Kuiken & Vedder (2017)	Dutch	Various	A2-B1	*N* = 32	*N* = 17	Decision		*N* = 4	X	
	Italian	Dutch	A2-B1	*N* = 39	*N* = 18			*N* = 4		
Kuiken & Vedder (2018)	Dutch	Various	A2-B2	*N* = 22		Decision		*N* = 4		X
	Italian	Dutch	A2-C1	*N* = 26				*N* = 4		

Appendix C. Studies in which the FA rating scale has been used

Studies (in chronological order)	Language						Raters		Modality	
	NNS									
	L2	L1	Level	*N*	NS	Task type	Expert	Nonexpert	Writing	Speaking
Révész, Ekiert & Torgersen (2016)	English	Japanese Spanish	Low-interm. Intermediate Low-adv. Advanced	*N* = 20 *N* = 20 *N* = 20 *N* = 20	*N* = 20	Complaint Refusal Narration Advice Summary	*N* = 10	*N* = 10		X
Faone & Pagliara (2017)	Italian	Chinese	A2-B1	*N* = 15		Instruction Narration	*N* = 3	*N* = 3	X	
Pallotti (2017a, b, c, this issue)	Italian	Various	Grade 3–5	*N* = 64	*N* = 153	Narration		*N* = 10	X	
Ekiert, Lampropoulou, Révész & Torgersen (2018)	English	Japanese Spanish	Low-interm Intermediate Low-adv. Advanced	*N* = 20 *N* = 20 *N* = 20 *N* = 20	*N* = 20	Complaint Refusal Advice	*N* = 2			X
Herraiz Martínez (2018); Herraiz Martínez & Alcón Soler (2019)	English	Spanish Catalan	A2-C1	*N* = 102		Motivation letter	*N* = 3		X	
Del Bono (2019, 2020)	Italian	Dutch	A2-B2	*N* = 15		Decision Narration Instruction		*N* = 5	X	
De Meo, Maffia & Vitale (2019)	Italian	Various	A2	*N* = 50		Description Narration Interaction	*N* = 2		X	

Studies (in chronological order)	Language						Raters		Modality	
	NNS									
	L2	L1	Level	*N*	NS	Task type	Expert	Nonexpert	Writing	Speaking
Orrù (2019)	Italian	Hungarian	A2-C1	*N* = 40		Decision Narration Instruction		*N* = 4	X	
Orrù & Foti (2020)	Italian	Hungarian	A2-C1	*N* = 40		Decision		*N* = 4	X	
Nuzzo & Bove (2020)	Italian	Various	(Low-)interm. to (low-)adv.	*N* = 20	*N* = 20	Decision Narration Instruction		*N* = 7	X	
Ekiert, Révész, Torgersen & Moss (this issue)	English	Spanish	Low-interm. Intermediate Low-adv. Advanced	*N* = 10 *N* = 10 *N* = 10 *N* = 10		Complaint Refusal	*N* = 2			X
Nuzzo & Bove (this issue)		Italian	Native speakers		*N* = 30	Motivation letter	*N* = 3	*N* = 15	X	
Strobl & Baten (this issue)	German	Dutch	B2	*N* = 30		Narration	*N* = 3		X	
Martín Laguna (in preparation)	English	Spanish	Beginners to advanced	*N* = 25 *N* = 25		Decision Narration Instruction		*N* = 8	X	

Address for correspondence

Folkert Kuiken
Amsterdam Center for Language and Communication (ACLC)
University of Amsterdam
Postbus 1637
Amsterdam 1000 BP
the Netherlands

f.kuiken@uva.nl

Biographical notes

Folkert Kuiken is Professor Emeritus of Dutch as a Second Language and Multilingualism at the University of Amsterdam, The Netherlands, and Academic Director of the Institute for Dutch Language Education at that same university. His research interests include the effect of task complexity and interaction on SLA, Focus on Form, and the relationship between linguistic complexity and functional adequacy. He (co)authored and (co)edited various books and special issues, including "Dimensions of L2 performance and proficiency" (Housen, Kuiken & Vedder, 2012).

https://orcid.org/0000-0002-1667-2501

Ineke Vedder is researcher at the University of Amsterdam, The Netherlands. Her research interests include Instructed Second Language Acquisition (particularly Italian), academic writing in L2 and L1, L2 pragmatics, and assessment of functional adequacy in L2 performance, in relation to linguistic complexity. Her publications have appeared in various edited books and journals, comprising two special issues, together with Housen, De Clercq and Kuiken, on syntactic complexity in SLA research (2019).


https://orcid.org/0000-0002-2677-0228

Publication history

Date received: 25 February 2021
Date accepted: 18 March 2022
Published online: 20 June 2022