10/05/2024 22:08

(Article begins on next page)

Holistic and Analytic Assessment of Functional Adequacy

Abstract

.
This study looks at the correlation between functional adequacy (FA), holistically assessed, and analytic linguistic measures, in a corpus of texts written by Italian monolingual and multilingual primary school pupils. Texts were first evaluated using the FA rating scales by Kuiken & Vedder (2017, 2018), plus one for Coherence & Cohesions from the CEFR (Council of Europe, 2001'). They were then coded for a number of features directly bearing on FA and its subdimensions. Results show correlations between holistic scores and analytic measures, such as thost between Content and the number of words or secondary idea units (r = .59 / .65). Others were less strong, yet going in the expected direction, e.g. more ambiguous referential expressions were negatively correlated to Comprehensibility. Correlations were generally stronger for monolingual than for multilingual children.

Keywords: functional adequacy, writing, text quality measures, primary school pupils, assessment

Providing detailed and valid descriptions of linguistic performance in communicative tasks has been one of the key aims of research on task-based language teaching (TBLT) and task-based language assessment (TBLA). The main methodological options may be classified according to two main dimensions. The first is whether the assessment concerns the structure or the function of a linguistic text. If the focus is on structure, the aim is to describe 'how the text is', its features on various levels, such as lexis, grammar, phonology, intonation. If the focus is on function, the aim is to describe 'how the text works', its impact on an external world of actors and actions, its role in a communicative situation or task. The second dimension concerns how the assessment is carried out, whether by human raters who evaluate the text holistically, according to their impressions, possibly guided by descriptor scales, or by coding and counting specific phenomena. Some of these measures can be obtained automatically, such as various fluency indicators, text length, syntactic complexity, lexical diversity and sophistication; a few of these analytic counts, however, do require a degree of human interpretation, as is the case with some accuracy scores.

The relationships among these dimensions are represented in Table 1. Cell 1 is exemplified by many rating scales commonly employed in language testing and assessment (LTA); their use in second language acquisition (SLA) research is more limited, although not completely absent (see for instance Kuiken et al., 2010). Rating scales in Cell 2 are also frequently included in language tests, where they complement those in Cell 1, by adding a communicative dimension to the evaluation; the Functional Adequacy scales proposed by Kuiken and Vedder (2017) are an example from SLA research. The analytic measures of Cell 3, including but not limited to the complexity, accuracy and fluency (CAF) triad, have been extensively employed in SLA studies, where they have long been the main or sole focus of interest; in language testing, they have been used mainly to validate holistic ratings, as for instance by Iwashita et al. (2008). Finally, Cell 4 contains some indications of how functional

adequacy (henceforth, FA), or communicative effectiveness, may be operationalized with objective measures, although this approach is not so common in either SLA or LTA.

Table 1
Describing linguistic performance in communicative tasks

	Structure	Function
Holistic-subjective	1. Rating scales of accuracy, complexity, fluency and other linguistic aspects.	2. Rating scales of adequacy, appropriateness, effectiveness, etc.
Analytic-objective	3. Measures of CAF, textual connectives, type of pronouns, grammatical processes, etc.	4. Number of correctly selected items in jigsaw task, path followed in map task, other objective measures of successful task completion.

This article investigates the relationships between FA holistic ratings (Cell 2) and analytic measures of text structure (Cell 3), in an attempt to clarify whether some linguistic features are associated to raters' perception of FA and task fulfilment.

Previous Research
The first LTA studies on the relationship between holistic ratings and analytic measures of linguistic performance date back to the 1990's. For instance, Douglas (1994) compared test scores with various aspects of test-takers' speech, noting that there were frequent discrepancies between the two modes of assessment. Fulcher (1996) analyzed the transcripts of 21 oral interviews, finding that a number of aspects of fluency, such as pause location or types of hesitations, could predict how the same construct was scored by human raters. Iwashita et al. (2008) conducted a large-scale study showing that holistic scores in a speaking test were related to various linguistic features in the candidates' productions, such as grammatical accuracy, lexical richness, target-like pronunciation, speech rate, while the effect of grammatical complexity was much smaller. In subsequent years the number of studies investigating the relationship between test scores and specific linguistic features has constantly grown, and this type of analysis has become a standard tool for analyzing and validating language assessments (for recent contributions, see Gu & Hsieh, 2019; Khushik & Huhta, 2020; Lahuerta Martínez, 2018 and references therein).

While in LTA holistic ratings were long established before the introduction of more analytic measures, the opposite occurred in SLA research. From its very beginnings, with error analysis in the 1960's, interlanguage analysis since the 1970's and the research program on complexity, accuracy and fluency begun in the 1990's, the focus has always been on describing L2 productions analytically, by looking at specific linguistic features (Cell 3 in Table 1). Typical research questions concerned how different dimensions of the CAF triad evolved over time or varied across task conditions, but the fundamental questions 'Is the message adequate and effective? Does it achieve its goals?' were seldom asked. Pallotti (2009) was one of the first to notice this paradox. While most research on CAF was based on communicative tasks, analyses normally concentrated on whether a task elicited longer or shorter clauses, more or less varied vocabulary, faster or slower speech, but not on whether the task's extralinguistic goals were fulfilled (e.g. placing objects on a map, solving a problem, persuading or informing an audience), that is, whether communication was adequate (Cell 4).

In the following years some SLA researchers began to include FA in their investigations. For instance, Kuiken et al. (2010) rated their participants' written compositions using two holistic scales, one for linguistic complexity, the other for communicative adequacy, both inspired by descriptor scales from the Common European Framework of Reference (CEFR, Council of Europe, 2001) and the proficiency scales developed in the WISP project (What Is

Speaking Proficiency; De Jong et al., 2012). Both studies found that accuracy and lexical diversity were strong predictors of adequacy ratings, while syntactic complexity had little or no effects. Similar results were obtained by Hulstijn et al. (2012), who found that analytic measures of grammatical accuracy and lexical diversity correlated with CEFR levels, while syntactic complexity played a small role and only at higher proficiency levels.

Révész et al. (2016) looked at whether the relationship between FA and linguistic features varies across different oral tasks, using both task-dependent and task-independent rating scales and an array of CAF measures. They found that fluency was the strongest predictor of "communicative adequacy" (a construct akin to FA), with other dimensions, such as lexical diversity, grammatical and connector accuracy, and syntactic complexity also playing a role, albeit smaller.

All these studies involved adult participants. Research on children normally does not mention notions such as communicative/functional adequacy, although a number of studies were carried out to investigate the relationships between overall writing quality (assessed with holistic ratings) and more specific analytic features. Vocabulary has been shown to play a major role in predicting writing quality judgments, in terms of both the variety and sophistication of lexical items (Durrant & Durrant, 2022; Olinghouse & Wilson, 2013). Handwriting fluency is also an important factor, especially in grades 3 and 4, when children may have reached quite different levels of automatization of this fundamental skill (Roessingh et al., 2019; Skar et al 2021). This shows that predictors of writing quality may vary between children and adults, as the latter for instance are supposed to have all reached a high and stable degree of handwriting fluency.

Within the field of SLA research, Kuiken and Vedder (2017) proposed an operationalization of the FA construct along four dimensions, Content, Task Requirements, Comprehensibility, Coherence & Cohesion (see Kuiken and Vedder, this issue, for a more comprehensive presentation), testing its applicability on written argumentative texts by native speakers and L2 learners of Dutch and Italian. They found that there was a good level of interrater agreement in applying the scales, that results of L1 and L2 speakers were clearly differentiated, and that the four subdimensions were highly correlated, with *r* values ranging from .544 to .938.

While these scales offer a valuable contribution to a better definition and operationalization of the FA construct, some unresolved issues remain. The first is whether the four subscales may be added up to form a unitary FA score. Kuiken and Vedder (2017) refer to a rating 'scale', but results are always given independently for the four subscales. A second issue has to do with the definition of these subconstructs. Some of them are composite, beginning with Coherence & Cohesion, which, as the very name suggests, concerns two different aspects. While this conjoined phrase is widespread in the teaching and assessment literature, the two notions are theoretically distinct and there may be empirical cases where one is present and the other absent, as with texts with a very coherent and logical flow of ideas, but little or no cohesive devices, or texts with an intricate network of cohesive links superimposed on a very confused conceptual structure. This also occurs with the scale for Content, whose descriptors refer to both the quantity of ideas and to their being 'consistent with' or 'unrelated to' each other, which introduces a dimension of coherence in this scale, too. Finally, the status of the Task Requirements subscale is unclear, as it seems to correspond to FA as a whole – if all the task's requirements are satisfied, then performance should be deemed as functionally adequate. Furthermore, all the descriptors in this scale ask whether the task's 'questions' have been 'answered', which may be relevant for some but not all communicative situations.

This article will investigate the relationships between Kuiken and Vedder's (2017) FA scales, supplemented by the CEFR scale for coherence and cohesion, and a series of linguistic features. Most of these features, however, do not come from the standard set of CAF measures, but represent aspects that have more directly to do with FA. In fact, CAF and FA represent independent dimensions – some texts may be very complex, accurate and fluent, yet not reach a communicative goal, and vice versa, at least from a theoretical point of view. In practice, as the

studies reviewed above show, there may be some empirical correlations between FA scores and CAF measures, that tend to be stronger as regards fluency, less so for accuracy, with a rather weak relationship to (syntactic) complexity. This stands to reason, given that the last aspect has more to do with stylistic preferences and is less essential for task fulfilment. The FA rating scales proposed by Révész et al. (2016) and Kuiken and Vedder (2017) do not mention complexity, accuracy and fluency, which is coherent with the idea that these constructs should be treated independently. However, as Révész et al. (2016) suggested, there might be some more specific measures directly tapping into the FA construct, or some of its subdimensions, and the present study will develop this point. Furthermore, it will investigate how these analytic measures correlate not only with FA as a whole, but also with its subdimensions of Comprehensibility, Content and Coherence & Cohesion, looking at children's data, which have not been the focus of previous studies.

Methodology

The study will seek to answer the following research questions:
1   To what extent are analytic measures of text quality related to holistic FA ratings?
2   To what extent are analytic measures of text quality related to the FA subdimensions of Content, Comprehensibility and Coherence & Cohesion?
3   What are the correlations among different subdimensions of FA and with FA as a whole?
4   How is Kuiken and Vedder's (2017) rating scale on Coherence & Cohesion related to a similar scale from the CEFR?

This study is mainly exploratory, investigating how holistic ratings of FA are related to a number of analytic measures. The aim is thus not to test a pre-determined theoretical model, but to establish what variables best predict FA scores, in order to stimulate further research and contribute to improving the scales and analytic measures related to them. Data consist in texts written by pupils in grade 3, 4 and 5 (age 8-10) in Italian primary schools), further subdivided into monolinguals and multilinguals. The term 'multilingual' refers to children who employ one or more languages other than Italian in their daily communicative activities, and thus does not apply to those whose additional language is just a little English learned and used at school. Inclusion in the multilingual group was based on children's self-declarations together with teachers' reports. While many of these multilingual children learned Italian as an additional language after acquiring the family language, and may thus be considered L2 learners, there were also several cases for which it was difficult to draw a clear boundary between L1 and L2 (or Ln) acquisition, such as various forms of simultaneous multilingualism with different ages of onset for different languages, or complex patterns of differential language competence and use. For these reasons, in this particular context the monolingual-multilingual distinction was deemed to be more appropriate than others, like L1/L2 users or native/non-native speakers. In the total sample (N = 217), there were 153 monolinguals and 64 multilinguals.

Children watched twice a five minutes video clip from a silent movie of the 1930's by Harold Lloyd, without taking notes. They were told that they would receive no grades, but that they should have strived to tell the story to the best of their capacities to a teacher who had not seen it, which was indeed the case, making the activity an authentic communicative task. After that, they wrote spontaneous texts with no time limits, and they generally completed the task in about 30-50 minutes, depending on age and individual writing skills.

The hand-written texts were transcribed without any editing and they were made anonymous in order to ensure that subsequent ratings and codings could not be biased by knowing the authors' age, whether they were monolingual or multilingual or belonged to an experimental or control class. The texts were first rated according to three of the four FA scales from Kuiken and Vedder (2017), translated into Italian, with scores ranging from 1 to 6 (Content; Comprehensibility; Coherence & Cohesion). The scale on Task Requirements was not included as the task consisted in telling a story to a teacher who had not seen the film before, so that she could understand what happened; producing a clear and sufficiently detailed text

already fulfilled this requirement, and these two dimensions are already included in the Content and Comprehensibility scales. An additional scale on Coherence & Cohesion, with eight levels, was taken from the Common European Framework of Reference for Languages (Council of Europe, 2001), in order to compare these two instruments. Ratings were done by members of the research group and four students graduating in educational linguistics, who received no preliminary training on the rating scales but were provided with a detailed coding manual; 10% of the scripts were double-rated, with an average Intraclass correlation coefficient (two-way mixed effects model) of .85 (R core team, 2020, package irr).

Data Analysis

Texts were coded for several linguistic features which represented the subconstructs of the FA scales proposed by Kuiken and Vedder (2017)[1]. Coding was performed by researchers or students graduating in educational linguistics; in the latter case, analysis was double-checked and corrected by members of the research group. A 30-page coding manual was produced to ensure transparency and reliability of analyses and the few remaining dubious cases where adjudicated by the principal investigator.

The analytic measures considered in the current study were the following (with abbreviations used in the tables), grouped according to Kuiken and Vedder's (2017) subdimensions of FA.

Content
- Text length: total number of running words, or tokens (words)
- Number of main idea units (MIU; theoretical maximum = 5). For instance, the initial MIU in the first video was 'Charlot is hired as a night watchman'.
- Number of secondary idea units (SIU; theoretical maximum = 43). Examples of SIUs contained in the first MIU: 'people crowding outside the shopping mall', 'Charlot enters the shopping mall', 'Charlot is interviewed by the manager and is hired', 'the owner arrives and talks to Charlot and the manager' etc.

Main and secondary idea units were identified according to a very detailed list contained in the coding manual, describing virtually every action in the video clip.

Coherence & Cohesion
- Variety of textual connectives (connect), that is, inter-clausal linking expressions; variety was scored with Guiraud's (1954) Index of Lexical Richness to control for the effects of text length, i.e. dividing connectives' types by the square of connectives' tokens.[2]
- Commas per 100 words (comm). This measure indicates how the text is segmented into smaller syntactic and discursive units. Other punctuation marks were also scored, but won't be reported here for space limitations and because their correlation with FA dimensions was minimal (*r* always < .15).
- Inappropriate commas per 100 words (INcomm). As a complement to the previous measure, this parameter concerns the accuracy of comma use. Only clearly inappropriate uses were scored, as omissions proved to be too subjective a category for reliable analysis.
- Run-on sentences (RunOn). Texts by children and unexperienced writers often contain long, unorganized sentences, where ideas simply follow one another separated by commas or no punctuation at all, as in the following example: *When they went into the shoe department where Charlot saw some roller skates, he took them, then showed another pair to the girl and both wore them, Charlot ended first then he started making some exhibitions, then while he was skating backwards he went out of the department and got into a room where there was no railing and Charlot put on a blindfold and started making some exhibitions with the blindfold right next to the hole, but luckily without falling, the girl noticed it and rushed to save him*. A run-on sentence was operationalized as a stretch of text between two periods, colons or semicolons, containing any combination of eight or more different subjects and predicative units.[3]

5

- Inappropriate tense shifts per 100 words (T.Shifts). Verb shifts are rather common in primary school children's narrations (Kersten, 2009), and were defined as a change of verb tense that is not justified by textual reasons (e.g. the move from the description of background states to the narration of events in the plot), as in the following extract: *Harry <u>has</u> to take a train and he <u>saw</u> a lady and <u>lets</u> her pass first. The lady <u>starts</u> talking to the ticket controller and Harry <u>came</u> in quickly*.

Comprehensibility
- Lexical diversity (MATTR). A wider vocabulary range, reflected in more lexical diversity in the text, should lead to more adequate lexical choices, which in turn may have an impact on overall comprehensibility; on the contrary, it is difficult to get a clear and exhaustive message across with just a few lexemes. This dimension was measured with the Moving Average Type-Token Ratio (MATTR, Covington & McFall, 2010), consisting in calculating the TTR on equal-sized samples; sample size was set at 50 words, in order to include even the shortest texts (Zenker & Kyle, 2021). Given that Italian is a richly inflected language, MATTR was computed on lemmas rather than on inflected words.
- Ambiguous references to entities per 100 words: introductions (AmbInt), maintenances (AmbMan), reintroductions (AmbRei). In a narrative text, an entity can be introduced, upon its first mention, and then maintained in the following clauses. This referential chain may be interrupted with the introduction of new entities, but the original entity can be later on reintroduced in the text. Sometimes these references are ambiguous, in that they do not allow the reader to clearly identify the entity in question. For instance, a text beginning with *She bought a new dress* leaves the referent of *she* underspecified; similarly, the *he* in *John and his dad went to the park. He looked very tired* is ambiguous, as it may refer to either John or the father (a more thorough description of these categories is provided in Author 2021 and in the coding manual).

In the following pages, correlation and regression analyses will be reported. They are all based on linear models, for the following reasons. Global FA scores, resulting from the mean of three subscales from Kuiken and Vedder (2017), form a continuous range of values. Although scores on these three subscales and the one from the CEFR are values on six- or eight-point ordinal scales, using linear statistics is appropriate in this case, too, as their distribution does not strongly deviate from normality (skewness and kurtosis always < 1.0) and their relationships with predictor variables is always monotonic (Norman, 2010; Larson- Hall, 2016).

Results: Correlation Analysis
The correlation matrices for monolingual and multilingual pupils are reported in Tables 2 and 3). Shading in the first column represents how variables may be grouped according to the FA subscale they are more directly relevant for. The first five lines represent scores for the holistic rating scales, which are the dependent variables: Content, Comprehensibility (Compr), Coherence & Cohesion (co.coKV) as defined by Kuiken and Vedder (2017), plus Functional Adequacy (FA), which is the average of these three scores; the fifth line reports scores on the CEFR scale for Coherence & Cohesion (co.coCEF). The following variables are those theoretically related to the Content subdimension: text length expressed as the number of words (words) and the number of main (MIU) and secondary idea units (SIU); to Coherence & Cohesion: variety of connective devices (connect), commas per 100 words (comm), inappropriate commas per 100 words (INcomm), percentage of words in run-on sentences (RunOn), tense shifts per 100 words (T.Shifts); and to Comprehensibility: lexical diversity calculated with MATTR and referential ambiguity in introductions (AmbInt), maintenances (AmbMan) and reintroductions AmbRei).

Table 2 Correlation matrix - monolingual pupils (N = 153). * = p < .05; ** = p < .01 (two-sided, Holm's method)

	FA	Content	Compr	co.coKV	co.coCEF	words	MIU	SIU	conn	comm	INcomm	RunOon	T.Sfhifts	MATTR	AmbInt	AmbMan	AmbRei
FA		0.87	0.9	0.9	0.82	0.61	0.39	0.62	0.46	0.32	0	0.01	-0.19	0.49	-0.11	-0.17	-0.31
Content	**		0.63	0.64	0.72	0.59	0.46	0.65	0.37	0.22	-0.06	0.11	-0.16	0.34	-0.08	-0.09	-0.28
Compr	**	**		0.8	0.66	0.47	0.31	0.47	0.43	0.32	0.09	-0.07	-0.2	0.45	-0.15	-0.21	-0.22
Co.coKV	**	**	**		0.8	0.55	0.25	0.5	0.44	0.33	0	-0.02	-0.16	0.53	-0.08	-0.17	-0.32
o.coCEF	**	**	**	**		0.57	0.33	0.56	0.46	0.4	0.04	-0.01	-0.16	0.47	-0.03	-0.11	-0.3
words	**	**	**	**	**		0.47	0.89	0.31	0.38	0.01	0.26	-0.16	0.3	-0.14	-0.04	-0.25
MIU	**	**	*		**	**		0.54	0.17	0.24	0	0	-0.2	0.12	0.08	-0.11	-0.07
SIU	**	**	**	**	**	**	**		0.28	0.36	0	0.27	-0.2	0.27	-0.1	-0.02	-0.18
conn	**	**	**	**	**	**		*		0.22	-0.01	-0.08	-0.16	0.49	-0.13	-0.12	-0.23
comm	**		**		**	**		**			0.55	0.08	-0.21	0.36	-0.01	0.07	-0.1
INcomm										**		-0.03	-0.02	0.14	0.09	0.03	0.05
run.on													0.06	0.03	0	0.25	0.1
T.Shifts														-0.36	0.03	0.07	0.19
MATTR	**	**	**	**	**	**			**	**			**		-0.05	-0.05	-0.23
AmbInt																0.05	0.12
AmbMan																	0.48
AmbRei	**	*		**	*	*										**	

Table 3 Correlation matrix - multilingual pupils (N = 64)

	FA	Content	Compr	co.coKV	co.coCEF	words	MIU	SIU	conn	comm	INcomm	RunOn	T.Shifts	MATTR	AmbInt	AmbMan	AmbRei
FA		0.82	0.85	0.92	0.76	0.43	0.21	0.43	0.29	0.3	0.16	-0.08	-0.18	0.38	-0.04	-0.26	-0.25
Content	**		0.47	0.63	0.52	0.54	0.39	0.58	0.22	0.17	0.03	-0.14	-0.07	0.26	0	-0.21	-0.3
Compr	**	*		0.78	0.7	0.21	0	0.19	0.28	0.3	0.2	0.03	-0.24	0.41	-0.09	-0.2	-0.17
Co.coKV	**	**	**		0.78	0.34	0.13	0.32	0.26	0.3	0.2	-0.09	-0.18	0.32	-0.01	-0.26	-0.16
Co.coCEF	**	**	**	**		0.36	0.25	0.32	0.31	0.38	0.24	0.03	-0.3	0.29	-0.01	-0.17	-0.15
words	*	**		**	**		0.52	0.88	0.31	0.12	0.06	0.11	-0.01	0.12	0.11	-0.13	-0.35
MIU				*		**		0.58	0.13	0.24	0.08	0.04	0.07	0.13	0.16	0.11	-0.24
SIU	*	**		**	*	**	**		0.14	0.11	-0.03	0.02	0.02	0.15	0.07	-0.08	-0.31
conn				*	*	*				0.3	0.24	-0.05	-0.13	0.44	-0.03	-0.11	-0.31
comm				*	**			*			0.75	-0.11	-0.23	0.27	-0.12	0.06	-0.2
INcomm										**		0.01	-0.02	0.14	-0.15	0.02	-0.15
RunOn													-0.15	-0.14	0.14	0.13	0.07
T.Shifts					*									-0.17	0.09	0.04	0.03
MATTR				*	*				**	*					-0.27	0.07	-0.12
AmbInt														*		0.01	0.18
AmbMan				*													0.11
AmbRei						**		*	*								

1

Monolingual speakers' scores on subdimensions, such as Content, Comprehensibility and Coherence & Cohesion, correlate strongly (.87-.90) with FA, which is the average of the three, and less so, but still considerably, with one another (.63-.80). A similar picture emerges for multilinguals, although the correlation between Content and other dimensions, especially Comprehensibility (.47) is somewhat lower, which shows that some pupils in this group wrote texts perceived to be rich with ideas but not so cohesive or easy to understand. The correlation between the two scales on Coherence & Cohesion (Kuiken and Vedder's and the CEFR) is also rather high, for both monolingual and multilingual pupils (.80 and .78, respectively).

In monolinguals, FA and its subscales show moderate positive correlations (approximately between .4 and .6) with text length, the number of secondary idea units, lexical

diversity and connectives' variety. Weaker correlations (between .2 and .4) are obtained for main idea units and comma use and, negatively, for ambiguous reintroductions. The pattern is the same for multilingual pupils, although correlations tend to be weaker. Inappropriate commas, run-on sentences, tense shifts and ambiguous introductions and maintenances show virtually no correlations with FA and its subscales for either group. It should be noted that all correlations, regardless of their strength, go in the expected direction: text length, number of idea units, lexical diversity, connectives' variety and comma use, all typically regarded as valuable features in a written text, are positively related to FA and its subdimensions, while the relationship is negative for inappropriate commas, run-on sentences, tense shifts and ambiguous references, which denote problematic aspects.

Correlations among predictor variables (other than FA subdimensions) are in general rather weak, under .40, and the few exceptions are easily interpretable. The strongest is that between words and SIU (.89 and .88 in monolinguals and multilinguals), as texts expressing more ideas obviously tend to be longer. This is followed, again unsurprisingly, by that between commas and inappropriate commas (.55 and .75). MATTR is also positively related to connectives' variety (.49 and .44), given that the latter measure taps lexical diversity in a specific domain.[4]

Results: Regression Models
 In the following pages, two complementary approaches to regression analysis will be taken, bottom-up and top-down. In the bottom-up, exploratory, approach, all predictor variables are potentially relevant and no theoretical assumptions are made regarding their role in predicting the outcome variable. The 'relative importance' of each predictor variable will be calculated using the LMG method (Lindeman et al., 1980; for a review of relative importance metrics, see Grömping, 2006, 2015; applications to SLA research are discussed by Larson-Hall, 2016). This amounts to performing a series of hierarchical regressions, systematically changing the order of variables, and calculating the average squared semi-partial correlation (sr^2) value of each of them, which gives the amount of explained variance contributed by each factor; the relative importance of all factors added together amounts to the model's total explained variance (R^2). This method combines the advantages of hierarchical regression, in which all the sr^2 values add up to the model's total R^2, with that of standard regression, that does not require one to choose a theoretically motivated variable order, which, in the current state of knowledge, would not have sufficient empirical grounds. The top-down approach, on the other hand, will be implemented by proposing selective regression models including only the theoretically relevant variables for each subscale.

The high correlation between the predictor variables 'words' and secondary idea units (SIU) was resolved by discarding the variable 'words'. The reasons for this choice are firstly logical: there is an asymmetrical causal relation between the two variables, given that expressing more ideas necessarily requires more words, but the reverse is not always true: a text may be long because it is uselessly verbose, or because it contains off-topic information that does not contribute to task fulfilment. This insight was also supported by empirical observations of a simple model with just 'words' and SIU as predictors and Content as the dependent variable, which is the subdimension conceptually and empirically more related to text length and the number of idea units. Here, adding SIU to a monovariate model with 'words' as unique predictor, increases R^2 by .076 in monolinguals and .049 in multilinguals, whereas adding 'words' to SIU has virtually no effect on the model's explained variance (+ .000 and + .005, respectively). This also provides evidence to the fact that, when applying the Content scale, raters paid more attention to the number of ideas than to sheer text length.

In order to reduce the number of predictors and thus increase the reliability of the regression models, run-on sentences and inappropriate commas were also excluded, given their very low correlations with FA and any of its subdimensions. In fact, regression models including these two variables explained less than 1% additional variance compared to models without them, and their effect could thus be considered to be negligible.

The assumptions for linear regression (Larson-Hall, 2016) were met in all models. For both monolinguals and multilinguals, multicollinearity was relatively low (variance factor, always < 2.0), as well as the presence of influential outliers (Cook's distance always < .05). Visual inspection of Residual vs. Fitted and Normal Q-Q plots showed there were no major issues with heteroscedasticity or normality of residuals.

Functional Adequacy

We will begin by looking at how all the predictor variables had an impact on FA, operationalized as the average of scores from the three subscales of Content, Comprehensibility and Coherence & Cohesion (Kuiken and Vedder 2017). Results for monolingual and multilingual pupils are given in Tables 4 and 5, respectively. The table's columns first report the relative importance metric (Relimp), followed by the regression coefficients (B) and their standard error ($SE\ B$), standardized coefficients (β), t statistics and p value. The total variance explained by the model (R^2, corresponding to the sum of Relimp values) is 55.4% for monolinguals and 40.3% for multilinguals.

Table 4
Functional adequacy - monolingual pupils

	Relimp	*B*	*SE B*	β	t	*p*
MIU	0.06	0.16	0.11	0.10	1.44	0.15
SIU	0.21	0.06	0.01	0.44	6.04	0.00
conn	0.08	0.33	0.13	0.17	2.50	0.01
comm	0.03	0.01	0.02	0.01	0.22	0.83
T.Shifts	0.01	0.03	0.03	0.06	0.99	0.32
MATTR	0.11	5.22	1.38	0.27	3.78	0.00
AmbInt	0.01	-0.10	0.21	-0.03	-0.49	0.62
AmbMan	0.01	-0.08	0.07	-0.07	-1.12	0.27
AmbRei	0.03	-0.10	0.07	-0.10	-1.42	0.16

Table 5
Functional adequacy - multilingual pupils

	Relimp	*B*	*SE B*	β	t	*p*
MIU	0.02	-0.12	0.19	-0.09	-0.62	0.54
SIU	0.12	0.05	0.02	0.37	2.75	0.01
conn	0.03	0.02	0.29	0.01	0.07	0.94
comm	0.04	0.07	0.05	0.19	1.58	0.12
T.Shifts	0.02	-0.03	0.03	-0.08	-0.77	0.45
MATTR	0.09	4.78	2.05	0.30	2.33	0.02
AmbInt	0.00	0.55	0.86	0.07	0.63	0.53
AmbMan	0.06	-0.24	0.11	-0.24	-2.19	0.03
AmbRei	0.02	-0.06	0.12	-0.06	-0.52	0.60

Results are similar for the two groups. The strongest predictors, in terms of relative importance and standardized coefficients, are secondary idea units (SIU) and lexical diversity (MATTR), which are also statistically significant. Connectives' variety plays a bigger role for monolinguals, while ambiguous maintenances (AmbMan) seem to be a better predictor of FA for multilinguals. All other predictors explain a smaller amount of the total variance, between 2% and 4% each. However, the coefficients for all of them (with a few exceptions whose value

is very close to 0) go in the expected direction, that is, they are negative for ambiguous references and unjustified tense shifts, and positive for all the other predictors.

Content

The first subdimension to be considered is Content, which has less to do with specific linguistic features and more with semantics and quantity and quality of information. The exploratory model including all predictors (Tables 6 and 7) explained 52.6% of variance for monolingual pupils and 41.7% for multilinguals. The most important factors, quite expectedly, were main and secondary idea units, accounting for 10% and 26% of total variance in monolinguals and 7% and 21% in multilinguals. A much smaller role was played, in the monolinguals' sample, by connectives' variety and lexical diversity; the latter also explained about 4% of total variance in the multilingual group, but here connectives had virtually no importance, while the presence of ambiguous references seemed to have a larger impact on Content ratings.

Table 6
Content - all predictors. Monolingual pupils

	Relimp	B	$SE\ B$	β	t	p
MIU	0.10	0.40	0.15	0.19	2.64	0.01
SIU	0.26	0.09	0.01	0.49	6.56	0.00
conn	0.05	0.35	0.18	0.13	1.94	0.05
comm	0.01	-0.05	0.03	-0.10	-1.44	0.15
T.Shifts	0.01	0.03	0.04	0.05	0.78	0.44
MATTR	0.04	3.56	1.89	0.14	1.88	0.06
AmbInt	0.00	-0.03	0.29	-0.01	-0.10	0.92
AmbMan	0.00	0.06	0.10	0.04	0.65	0.52
AmbRei	0.03	-0.21	0.10	-0.15	-2.17	0.03

Table 7
Content - all predictors. Multilingual pupils

	Relimp	B	$SE\ B$	β	t	p
MIU	0.07	0.17	0.24	0.10	0.70	0.49
SIU	0.21	0.07	0.02	0.45	3.34	0.00
conn	0.01	0.01	0.35	0.00	0.03	0.98
comm	0.01	0.02	0.06	0.04	0.31	0.75
T.Shifts	0.00	-0.01	0.04	-0.04	-0.33	0.74
MATTR	0.04	3.45	2.53	0.17	1.37	0.18
AmbInt	0.00	0.25	1.06	0.03	0.24	0.81
AmbMan	0.04	-0.23	0.14	-0.18	-1.70	0.10
AmbRei	0.04	-0.11	0.14	-0.09	-0.79	0.43

A more parsimonious model, presented in Tables 8 and 9, includes only those predictors that are conceptually related to the Content construct, that is, main and secondary idea units. The total variance explained by this model was 44.3% for monolingual and 33.9% for multilingual pupils, which shows that these two predictors alone are a satisfactory base for predicting scores on the Content rating scale.

Table 8

Content - selected predictors. Monolingual pupils

	Relimp	*B*	*SE B*	β	t	*p*
MIU	0.12	0.34	0.15	0.16	2.25	0.03
SIU	0.33	0.10	0.01	0.56	7.82	0.00

Table 9

Content - selected predictors. Multilingual pupils

	Relimp	*B*	*SE B*	β	t	*p*
MIU	0.08	0.14	0.22	0.08	0.65	0.52
SIU	0.26	0.08	0.02	0.53	4.13	0.00

Comprehensibility

Scores on the comprehensibility scale could not be easily predicted by any set of variables. The model with all predictors explained 41.2% of the variance for monolinguals and 32.3% for multilinguals (Tables 10 and 11). The variable with the highest importance is lexical diversity, accounting for 10% and 11% of total variance for monolinguals and multilinguals, respectively. Connectives' variety plays a smaller role for monolinguals (8%), and even smaller for multilinguals (3%), for whom comma use has a slightly higher predictive power (6%). Ambiguous references, that should have been in principle more directly related to comprehensibility, don't seem to predict much variance for either monolinguals and multilinguals.

Table 10

Comprehensibility - all predictors. Monolingual pupils

	Relimp	*B*	*SE B*	β	t	*p*
MIU	0.03	0.11	0.13	0.06	0.81	0.42
SIU	0.11	0.04	0.01	0.28	3.44	0.00
conn	0.08	0.37	0.16	0.17	2.30	0.02
comm	0.04	0.04	0.03	0.09	1.30	0.20
T.Shifts	0.01	0.01	0.04	0.01	0.18	0.86
MATTR	0.10	5.08	1.66	0.25	3.06	0.00
AmbInt	0.01	-0.31	0.26	-0.08	-1.21	0.23
AmbMan	0.03	-0.21	0.09	-0.18	-2.41	0.02
AmbRei	0.01	0.04	0.08	0.03	0.44	0.66

Table 11

Comprehensibility - all predictors. Multilingual pupils

	Relimp	B	$SE\ B$	β	t	p
MIU	0.02	-0.35	0.23	-0.22	-1.52	0.14
SIU	0.03	0.03	0.02	0.21	1.45	0.15
conn	0.03	0.01	0.34	0.00	0.02	0.99
comm	0.06	0.10	0.06	0.23	1.79	0.08
T.Shifts	0.03	-0.04	0.04	-0.11	-0.97	0.34
MATTR	0.11	6.41	2.47	0.35	2.60	0.01
AmbInt	0.00	0.66	1.04	0.08	0.63	0.53
AmbMan	0.04	-0.22	0.13	-0.19	-1.62	0.11
AmbRei	0.01	-0.07	0.14	-0.06	-0.49	0.63

This is confirmed in the more selective models where only theoretically relevant variables are included. These, in the case of Comprehensibility, are lexical diversity (using a more varied lexicon should produce a more precise message) and ambiguous references. Results in Tables 12 and 13 show that, when other predictors are excluded, lexical diversity remains by far the most important variable related to comprehensibility, but the presence of ambiguous references does not seem to have any special role, except perhaps for ambiguous maintenances, which compromise referential continuity and thus information flow. The total variance explained by these restricted models is 32.3% for monolinguals and 23.2% for multilinguals.

Table 12
Comprehensibility - selected predictors. Monolingual pupils

	Relimp	B	$SE\ B$	β	t	p
MATTR	0.19	8.84	1.49	0.43	5.92	0.00
AmbInt	0.02	-0.43	0.28	-0.11	-1.57	0.12
AmbMan	0.03	-0.19	0.09	-0.17	-2.04	0.04
AmbRei	0.02	-0.04	0.09	-0.03	-0.38	0.70

Table 13
Comprehensibility - selected predictors. Multilingual pupils

	Relimp	B	$SE\ B$	β	t	p
MATTR	0.16	7.71	2.17	0.42	3.55	0.00
AmbInt	0.00	0.38	1.01	0.04	0.37	0.71
AmbMan	0.04	-0.25	0.13	-0.22	-1.91	0.06
AmbRei	0.02	-0.12	0.13	-0.10	-0.90	0.37

Coherence & Cohesion – Kuiken and Vedder

Text coherence and cohesion assessed with the scale by Kuiken and Vedder (2017) could be predicted, in the monolinguals' sample, mainly by the number of secondary idea units and lexical diversity (respectively 14% and 15% of total variance explained); the variety of textual connectives and comma use, which are conceptually more related to this dimension, account only for 7% and 4% of total variance, with a smaller role played by ambiguous reintroductions, that also have to do with referential continuity. Essentially the same factors are involved in the multilinguals' model, but here none of them stands out: secondary idea units, lexical diversity, comma use and ambiguous maintenances explain 6-7% of total variance each, with other variables having a very small impact, including connectives' variety (just 2% of total variance explained). Total R^2 is 0.480 for the monolinguals' model and 0.317 for multilinguals (see Tables 14 and 15).

Table 14
Coherence & Cohesion Kuiken and Vedder - all predictors. Monolingual pupils

	Relimp	*B*	*SE B*	β	t	*p*
MIU	0.02	-0.03	0.12	-0.02	-0.27	0.79
SIU	0.14	0.05	0.01	0.36	4.57	0.00
conn	0.07	0.28	0.15	0.13	1.88	0.06
comm	0.04	0.03	0.03	0.07	0.98	0.33
T.Shifts	0.01	0.05	0.03	0.10	1.54	0.13
MATTR	0.15	7.05	1.52	0.35	4.63	0.00
AmbInt	0.00	0.03	0.23	0.01	0.14	0.89
AmbMan	0.01	-0.10	0.08	-0.09	-1.22	0.23
AmbRei	0.04	-0.13	0.08	-0.12	-1.65	0.10

Table 15
Coherence & Cohesion Kuiken and Vedder - all predictors. Multilingual pupils

	Relimp	*B*	*SE B*	β	t	*p*
MIU	0.01	-0.17	0.23	-0.11	-0.77	0.45
SIU	0.07	0.04	0.02	0.30	2.04	0.05
conn	0.02	0.05	0.33	0.02	0.14	0.89
comm	0.06	0.10	0.05	0.24	1.87	0.07
T.Shifts	0.02	-0.02	0.04	-0.07	-0.63	0.53
MATTR	0.06	4.43	2.40	0.25	1.84	0.07
AmbInt	0.00	0.74	1.01	0.09	0.73	0.47
AmbMan	0.07	-0.28	0.13	-0.26	-2.18	0.03
AmbRei	0.01	0.00	0.14	0.00	-0.01	0.99

The short list of selected predictors included only those that specifically have to do with this dimension, that is, connectives' variety, comma use and the presence of unjustified tense shifts, which are interruptions of grammatical continuity. Ambiguous maintenances and reintroductions were also included, as they may represent what in the scale are described as 'unrelated progressions' and 'coherence breaks'. These more parsimonious models (Tables 16 and 17) obviously explain a smaller amount of variance (R^2 = 0.299 and 0.202), but are theoretically more motivated and logically more coherent: for instance, in the complete model for monolingual pupils, tense shifts had a medium-sized positive coefficient, while in the restricted model their coefficient is small but, more sensically, negative. Connectives' variety is a relatively strong predictor for monolingual participants, accounting for 14% of their scores' total variance, whereas this is considerably lower (4%) in multilinguals. Comma use is a statistically significant predictor for both monolinguals and multilinguals, while tense shifts have a very small negative effect on Coherence & Cohesion scores. Referential ambiguities also seem to play a role on this scale, although ambiguous reintroductions matter more for monolinguals and ambiguous maintenances for multilinguals.

Table 16
Coherence & Cohesion KV - selected predictors. Monolingual pupils

	Relimp	B	SE B	β	t	p
conn	0.14	0.68	0.15	0.33	4.55	0.00
comm	0.08	0.09	0.03	0.24	3.33	0.00
T.Shifts	0.01	-0.01	0.04	-0.01	-0.16	0.87
AmbMan	0.01	-0.06	0.09	-0.05	-0.68	0.50
AmbRei	0.06	-0.21	0.09	-0.19	-2.35	0.02

Table 17
Coherence & Cohesion KV - selected predictors. Multilingual pupils

	Relimp	B	SE B	β	t	p
conn	0.04	0.33	0.32	0.13	1.03	0.31
comm	0.07	0.11	0.05	0.25	1.97	0.05
T.Shifts	0.02	-0.03	0.04	-0.09	-0.75	0.45
AmbMan	0.07	-0.28	0.13	-0.26	-2.16	0.04
AmbRei	0.01	-0.04	0.13	-0.04	-0.33	0.74

Coherence & cohesion - CEFR

The picture for the Coherence & Cohesion scale from the CEFR (Tables 18 and 19) is not very different from that obtained with Kuiken and Vedder's scale, although the total explained variance is slightly higher ($R^2 = 0.497$ for monolinguals and 0.339 for multilinguals). Here, too, ratings for monolinguals' texts seem to be primarily related to the number of secondary information units, followed by connectives' variety. Lexical diversity plays a smaller role, and even smaller is that of comma use and ambiguous reintroductions, while tense shifts are almost irrelevant. Factors predicting variance in multilinguals' scores are more spread out, and not exactly the same. Commas, connectives and tense shifts, which more logically pertain to this dimension, have the highest relative importance values (between 4% and 8%). Main and secondary idea units, together with lexical diversity, also play a role, but it is not as large as with monolinguals.

Table 18
Coherence & Cohesion CEFR - all predictors. Monolingual pupils

	Relimp	B	SE B	β	t	p
MIU	0.04	0.10	0.18	0.04	0.53	0.60
SIU	0.16	0.08	0.02	0.37	4.81	0.00
conn	0.09	0.65	0.22	0.21	2.93	0.00
comm	0.06	0.08	0.04	0.15	2.16	0.03
T.Shifts	0.01	0.07	0.05	0.09	1.34	0.18
MATTR	0.09	6.50	2.30	0.21	2.82	0.01
AmbInt	0.00	0.34	0.35	0.06	0.97	0.33
AmbMan	0.00	-0.04	0.12	-0.02	-0.32	0.75
AmbRei	0.04	-0.22	0.12	-0.13	-1.86	0.07

Table 19
Coherence & Cohesion CEFR - all predictors. Multilingual pupils

	Relimp	B	SE B	β	t	p
MIU	0.03	0.15	0.34	0.07	0.46	0.65
SIU	0.06	0.04	0.03	0.22	1.56	0.13
conn	0.04	0.40	0.50	0.11	0.81	0.42
comm	0.08	0.16	0.08	0.25	1.97	0.05
T.Shifts	0.06	-0.11	0.06	-0.22	-1.90	0.06
MATTR	0.04	3.48	3.58	0.13	0.97	0.34
AmbInt	0.00	0.54	1.51	0.04	0.36	0.72
AmbMan	0.03	-0.29	0.19	-0.17	-1.50	0.14
AmbRei	0.01	0.09	0.20	0.05	0.43	0.67

The more restricted models including only theoretically relevant variables (Tables 20 and 21), too, explain a slightly larger amount of variance than their counterparts using Kuiken and Vedder's (2017) scale ($R^2 = 0.342$ for monolinguals and 0.255 for multilinguals). Monolinguals' scores are predicted by connectives' variety and comma use, followed by ambiguous reintroductions, with virtually no predictive value for tense shifts and ambiguous maintenances. The strongest predictor for multilinguals' ratings is comma use, but here the presence of tense shifts has some importance, too (6% of explained variance). Connectives' variety, ambiguous maintenances and ambiguous reintroductions are relatively less important.

Table 20
Coherence & Cohesion CEFR - selected predictors. Monolingual pupils

	Relimp	B	SE B	β	t	p
conn	0.16	1.11	0.22	0.35	4.96	0.00
comm	0.12	0.18	0.04	0.31	4.35	0.00
T.Shifts	0.01	0.00	0.05	0.00	-0.04	0.97
AmbMan	0.01	-0.01	0.13	0.00	-0.06	0.95
AmbRei	0.05	-0.30	0.13	-0.18	-2.34	0.02

Table 21
Coherence & Cohesion CEFR - selected predictors. Multilingual pupils

	Relimp	B	SE B	β	t	p
conn	0.05	0.65	0.46	0.17	1.40	0.17
comm	0.10	0.18	0.08	0.28	2.31	0.02
T.Shifts	0.06	-0.11	0.06	-0.21	-1.80	0.08
AmbMan	0.03	-0.27	0.19	-0.16	-1.42	0.16
AmbRei	0.01	-0.02	0.19	-0.01	-0.10	0.92

Summary of findings
The first research question asked to what extent analytic measures of text quality are related to FA holistic ratings. The answer is that the nine predictor variables could explain 55.4% of monolinguals' FA score variance, and 40.3% for multilinguals, with quantity of secondary idea units (informative detail) and lexical diversity playing the biggest role (taken together, they accounted for 32% and 21% of total variance in the two groups). Connectives' variety had more predictive value for the monolinguals' group, while ambiguous references were more important in determining multilinguals' FA score.

As regards specific subdimensions, Content was the one with the most clear-cut results. Number of main and secondary idea units were by far the most important predictors in the full

model, and a model including just them could account for 44.3% and 33.9% of variance, for monolinguals and multilinguals, respectively. Text length was also highly correlated to this dimension, although the sheer number of words offered virtually no additional explanatory power after informativeness was factored in. This shows that raters correctly paid more attention to text content rather than to its length, and proves that the holistic impression of 'exhaustiveness' can be reliably translated into the analytic measure 'number of idea units'.

Comprehensibility was less easy to predict with the variables employed in this study (R^2 was 0.412 and 0.323 for the full model and 0.323 and 0.232 for the partial model, for monolinguals and multilinguals, respectively). Lexical diversity was clearly related to this dimension, as a text with more varied vocabulary is probably more precise and thus easier to understand. The presence of ambiguous references also contributed, negatively, to explaining some score variance on this dimension, although range of textual connectives and comma use had a slightly higher predictive value.

Similar amounts of explained variance were found for Coherence & Cohesion, as measured with Kuiken and Vedder's (2017) rating scale: R^2 for monolinguals and multilinguals was 0.480 and 0.317 in the full model and 0.299 and 0.202 in the model containing only theoretically relevant predictors. Slightly better results were achieved with the CEFR rating scale, with R^2 values of 0.497 and 0.339 for the full model and 0.342 and 0.255 for the restricted model. Here, as with other dimensions, monolinguals' performance was accountable based on a few, theoretically expectable, predictive variables, while multilinguals' score seemed to depend on a wider range of factors. In any case, connectives' variety, comma use and the presence of ambiguities in the referential flow all played a role in the raters' evaluation. It is worth recalling that two measures that could in principle have been related to text cohesion, that is, inappropriate commas and run-on sentences, were excluded from the regression analysis because of their very low correlation with FA in general and with the more specific scales on Coherence & Cohesion.

Implications for TBLT and directions for future research

These findings point to a number of implications for TBLT and TBLA research.

The first is that maximum clarity is needed as regards construct definition and operationalization, which involves much conceptual work well before the investigation begins. The relationships between FA and analytic measures should not just be 'discovered' after the fact, but they should be built into the key construct from the start. Thus, when developing descriptors of FA one should already think of their analytic correlates, which would contribute to more rigorous and coherent theoretical definitions. For instance, if the construct 'content' is defined in terms of the number of ideas, then we expect the latter measure to be a good predictor of the former; similarly, if descriptors of 'coherence and cohesion' explicitly mention the variety of linking expressions, analytic measures of such variety should positively correlate with ratings on this dimension. On the other hand, measures for empirical studies should be selected based on their conceptual relevance to the functional constructs they are supposed to index: rather than calculating omnibus CAF measures and reporting what correlates the best with FA, one should select those measures that meaningfully represent certain functional constructs. These measures do not necessarily need to belong to the CAF triad, but may assess, as in the present study, more specific aspects, like referential ambiguity, connectives' variety or punctuation.

When spelling out these relationships, one should bear in mind the distinctions made in Table 1. At the level of textual-linguistic structure (first column), holistic ratings and analytic measures are different *methods* of assessment, that target the same construct. For instance, one may judge text A to be longer than text B based on one's holistic impressions, or one may count the number of words; likewise, one may 'feel' that the lexicon is more varied in text A than in text B, or lexical diversity may be objectively calculated using some form of Type/Token ratio. In these cases, the analytic measure may be used to validate holistic impressions of text structure. It should be stressed that 'analytic' does not always coincide with 'objective', as if

analytic assessment could in principle always be performed automatically. While this may be true for text length or lexical diversity, some analytic measures, like number of idea units or ambiguous introductions, require human judgement, which, in such cases, is not applied to the whole text, as in holistic rating, but analytically, to individual structures.

This human judgment becomes essential when assessing a text's function (second column in Table 1): only a human being can establish whether a text is comprehensible, coherent or fulfills the task's requirements. There may be analytic-structural correlates to this holistic-functional appraisal, so that it may turn out, for instance, that texts perceived to be more comprehensible are longer, contain a more varied vocabulary, or fewer tense shifts. But these are empirical findings, not necessary and defining conditions. Some analytic features may represent rather neatly aspects of a holistic, functional construct, as is the case with secondary idea units representing content exhaustiveness. In other cases, the relationship may be weaker, or nil, from both a logical and empirical point of view. In our study, for instance, inappropriate use of the comma, or the presence of run-on sentences, seemed to play virtually no role in raters' perception of any of FA's dimensions. This might be due to how the functional constructs were defined (rating scales do not explicitly mention punctuation or sentence construction), or to how analyitic constructs were operationalized (our definition of run-on sentence was rather strict and it might have included some sentences that did not sound so improper).[5] However, it is also possible that here and in other cases the assessment of some dimensions of functional adequacy may be inherently independent of any analytic dimension. For instance, coherence (unlike cohesion) is probably impossible to define in terms of specific linguistic features. If we take a coherent text and scramble its sentences, the result will be a text with exactly the same values on most analytic dimensions, such as CAF, lexical diversity, connectives' variety, comma use etc., but that will receive completely different scores in terms of coherence (and comprehensibility). Likewise, in a classic experiment, Bransford and Johnson (1972) produced some passages lacking one specific piece of information, which made them incomprehensible; providing that information made the text entirely sensical and easy to understand, even though nothing changed in terms of structural, analytic features.

This means that it is not only difficult, but in many cases also theoretically and logically impossible to find structural correlates of functional dimensions like comprehensibility and coherence. In other words, some aspects of FA and its subdimensions can only be assessed with holistic ratings, and there is no way of reducing them to the analytic counting of linguistic features (Fulcher, 2015 makes a similar argument regarding fluency). In such cases, it would be wrong to see analytic measurement as validating holistic judgments, or to judge the quality of these measures based on their correlation with holistic ratings, because the two assessment methods target different, independent constructs.

While these methodological recommendations hold for all research in this area, the specific findings obtained in this study are clearly limited not only as regards sample size, but also by its particular context. For instance, the quality of children's writing may be assessed along different lines from that of adults', and some factors that impact the former, such as handwriting skills, may play a much more limited role in an adult population. Likewise, this study only looked at narrative texts, although the relationships between analytic measures and functional dimensions may be different in other genres. Preliminary research has shown that the link between analytic measures and functional adequacy does not seem to vary across tasks and modalities (Kuiken & Vedder, 2018; Révesz et al., 2016), but more studies are needed to explore this important area, again bearing in mind possible differences between children and adults.

Another dimension that has not received much attention in previous research, and that this study tried to address, are the similarities and differences between monolingual and multilingual users (a distinction partly corresponding to those between native and non-native spakers, or L1 / Ln users, more commonly employed in SLA research). The two groups in this study behaved similarly on most dimensions, showing that the relationship between analytic features and holistic ratings is not particularly affected by this factor. One important and systematic difference, though, is that these relationships tended to be weaker in the

multilinguals' subsample. This might be due to a higher variability among multilinguals, as their coefficients of variation for most measures were slightly larger than monolinguals', although differences were not very sizeable and in some cases inter-individual variation among multilinguals was even lower. What is probably the case is that in assessing multilinguals' performance raters may be influenced by a wider range of factors, beginning with linguistic competence, that provide a larger contribution to score variance than specific measures of written text quality. This also appears in the fact that, while monolinguals' scores can be predicted by relatively few important factors, for multilinguals a wider array of explanatory variables is involved.

This leads us to consider how FA is connected to other more general notions, like communicative competence or language proficiency, two terms that are in turn closely related ("the development of language proficiency should be guided and evaluated by the learner's ability to communicate", Savignon, 2018, p. 1). Indeed, a communicatively competent, or linguistically proficient, person should be able to produce functionally adequate texts, almost by definition. Thus one may wonder whether there is any difference between assessing FA and communicative competence or linguistic proficiency, except for the obvious fact that FA is a property of texts while competence and proficiency pertain to the persons producing them. In many previous studies on the relationship between analytic measures and holistic ratings, the latter concerned 'language proficiency' (Biber et al., 2016), 'speaking proficiency' (Iwashita et al, 2008; Hulstijn et al., 2012) or 'writing proficiency' (Crossley, 2020; Lahuerta Martínez, 2018). A direction for future research may thus be to clarify how these constructs are related, by answering questions like: is FA a manifestation of communicative competence, so that by assessing one the other can be assessed, too? What is the role of language proficiency in a wider communicative competence and, more particularly, in producing functionally adequate texts? How are different dimensions of language proficiency related to different subdimensions of FA? These theoretical questions may be addressed with empirical analyses, too. For instance, the correlations among subdimensions of FA found by Kuiken and Vedder (2017), Révész et al. (2016), and in the present study may suggest that there is a single latent variable, such as 'language proficiency', explaining why constructs like content richness and comprehensibility, which in theory should be independent of one other, are in practice correlated.

The correlations among FA subdimensions may also receive another explanation. In the present study, like in many others, the same rater scored different subdimensions one after the other for the same text. This may have caused some sort of halo effect, whereby scores on different dimensions tend to converge towards similar values. This possible validity threat may be checked by having different raters scoring different dimensions, although this would make rater severity a variable to control for, or, even better, have the same rater assessing multiple texts for one dimension at a time, and scoring different dimensions without knowing how the others were previously assessed.

These remarks lead us to the relationship between SLA and LTA research (Bartning et al., 2010). Both fields have investigated how analytic measures are related to, or predict, holistic ratings, with LTA playing a leading role both chronologically and in terms of the number of studies performed. In the present study, two scales on coherence and cohesion were employed, one developed by two SLA researchers (Kuiken & Vedder, 2017), the other from a text originating from, and having a considerable impact on, LTA, such as the CEFR. Results from the two scales were quite similar, with the latter being just slightly more predictable by analytic measures. Given that there is no shortage of holistic rating scales in LTA, one wonders whether and why SLA researchers should start producing their own. What is their added value? To what extent do they capture aspects that are not included in LTA rating scales, and, if this were indeed the case, how could the two research lines complement one another? Questions like these will continue to feed the fruitful debate on the relationships between SLA and LTA research, with FA playing an important mediating role across fields and disciplines.

References

Author 2019
Author 2021

Bartning, I., Martin, M.., & Vedder, I. (Eds.). (2010). *Communicative proficiency and linguistic development: Intersections between SLA and language testing research*. European Second Language Association.

Biber, D., Gray, B., & Staples, S. (2016). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics*, *37*(5), 639–668. https://doi.org/10.1093/applin/amu059

Bransford, J. D., & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 717–726. https://doi.org/10.1016/S0022-5371(72)80006-9

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

Covington, M., & McFall, J. (2010). Cutting the Gordian knot: The Moving-Average Type–Token Ratio (MATTR). *Journal of Quantitative Linguistics*, *17*(2), 94–100. https://doi.org/DOI: 10.1080/09296171003643098

Crossley, S. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, *11*(3), 415–443. https://doi.org/10.17239/jowr-2020.11.03.01

De Jong, N., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). 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—Investigating complexity, accuracy and fluency in SLA* (pp. 121–142). John Benjamins. https://doi.org/10.1075/lllt.32.06jon

Douglas, D. (1994). Quantity and quality in speaking test performance. *Language Testing*, *11*(2), 125–144. https://doi.org/10.1177/026553229401100203

Durrant, P., & Durrant, A. (2022). Appropriateness as an aspect of lexical richness: What do quantitative measures tell us about children's writing? *Assessing Writing*, *51*, 100596. https://doi.org/10.1016/j.asw.2021.100596

Fox, J., & Bouchet-Valat, M. (2020). *Rcmdr: R Commander*. https://socialsciences.mcmaster.ca/jfox/Misc/Rcmdr/

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing*, *13*(2), 208–238. https://doi.org/10.1177/026553229601300205

Fulcher, G. (2015). *Re-examining language testing: A philosophical and social enquiry*. Routledge.

Grömping, U. (2006). Relative importance for linear regression in R: The Package relaimpo. *Journal of Statistical Software*, *17*(1), 1–27.

Grömping, U. (2015). Variable importance in regression models. *WIREs Computational Statistics*, *7*(2), 137–152. https://doi.org/10.1002/wics.1346

Gu, L., & Hsieh, C.-N. (2019). Distinguishing features of young English language learners' oral performance. *Language Assessment Quarterly*, *16*(2), 180–195. https://doi.org/10.1080/15434303.2019.1605518

Guiraud, P. (1954). *Les caractères statistiques du vocabulaire. Essai de méthodologie*. Presses Universitaires de France.

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

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, *29*(1), 24–49. https://doi.org/10.1093/applin/amm017

Kersten, K. (2009). *Verbal inflections in L2 child narratives*. Wissenschaftlicher Verlag Trier.

Khushik, G. A., & Huhta, A. (2020). Investigating syntactic complexity in EFL learners' writing across Common European Framework of Reference Levels A1, A2, and B1. *Applied Linguistics*, *41*(4), 506–532. https://doi.org/10.1093/applin/amy064

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. Kim (Eds.), *Task-Based Language Teaching* (pp. 266–285). John Benjamins. https://doi.org/10.1075/tblt.10.11kui

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.

Lahuerta Martínez, A. C. (2018). Analysis of syntactic complexity in secondary education EFL writers at different proficiency levels. *Assessing Writing*, *35*, 1–11. https://doi.org/10.1016/j.asw.2017.11.002

Larson-Hall, J. (2016). *A guide to doing statistics in second language research using SPSS and R, 2nd edition*. Routledge.

Lee, Y. Y., & Ventura, S. (2017). *Lindia. Automated Linear Regression Diagnostic*. https://cran.r-project.org/package=lindia

Lindeman, R. H., Merenda, P. F., & Gold, R. Z. (1980). *Introduction to bivariate and multivariate analysis*. Scott, Foresman and Company.

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education*, *15*(5), 625–632. https://doi.org/10.1007/s10459-010-9222-y

Olinghouse, N. G., & Wilson, J. (2013). The relationship between vocabulary and writing quality in three genres. *Reading and Writing*, *26*(1), 45–65. https://doi.org/10.1007/s11145-012-9392-5

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

Pallotti, G. (2017). Applying the interlanguage approach to language teaching. *International Review of Applied Linguistics in Language Teaching (IRAL)*, *55*(4), 393–412. https://doi.org/doi.org/10.1515/iral-2017-0145

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

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

Roessingh, H., Nordstokke, D., & Colp, M. (2019). Unlocking Academic Literacy in Grade 4: The Role of Handwriting. *Reading & Writing Quarterly*, *35*(2), 65–83. https://doi.org/10.1080/10573569.2018.1499160

Savignon, S. J. (2018). Communicative competence. In J. I. Liontas (Ed.), *The TESOL Encyclopedia of English Language Teaching* (pp. 1–7). John Wiley & Sons. https://doi.org/10.1002/9781118784235.eelt0047Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing*, *47*, 100505. https://doi.org/10.1016/j.asw.2020.100505

Skar, G. B., Lei, P.-W., Graham, S., Aasen, A. J., Johansen, M. B., & Kvistad, A. H. (2021). Handwriting fluency and the quality of primary grade students' writing. *Reading and Writing*. https://doi.org/10.1007/s11145-021-10185-y

1   Unlike for English, the tools currently available for automatic analyis of Italian texts are scarce and contain just a few measures. For examples, lexical sophistication can only be assessed in terms of broad frequency bands (2K, 4K, 6K and beyond), which are not sufficiently fine-grained for tracking lexical development in young children. Most measures used in this study were thus scored by hand, and their identification rested on theoretical criteria, rather than on statistical reasons, such as picking those yielding significant findings.

2   While Guiraud index offers only a partial correction to text length effects (Zenker & Kyle, 2021), it seems the most viable option in this case, where the very small number of connectives' tokens per text ($M = 14.5$) makes it impossible to use more sophisticated measures like MATTR, MTLD or HD-D.

3   This operationalization currently rests on face validity; research is in progress on a more rigorous validation of the measure and the underlying construct, since, to the best of my knowledge, no widely accepted operational definitions are available.

4   Statistical analyses were performed with the R statistical software (R Core Team, 2020), using the packages Relaimpo (Grömping, 2006), Lindia (Lee & Ventura, 2017) and R Commander (Fox & Bouchet -Valat, 2020).

5   In future studies raters may be interviewed to know what criteria they used to assign scores, what features they paid attention to, or where they had doubts and difficulties. This would also provide important information on how the descriptor scales may be improved.