Writing performance and time of exposure in EFL immersion learners: analysing complexity, accuracy, and fluency

Isabel Tejada-SánchezCarmen Pérez-Vidal

Monday 20th August, 2018

1 Introduction

Since the emergence of immersion programmes in Canada in the late 1960s (**LambertTucker197** much research has been carried out on the development of linguistic competence in an L2 in such settings (**GeneseeStanley1976**; **Genesee1978**; **Swain2000french**). The majority of these studies, mostly in favour of immersion, have also addressed the limitations of immersion programmes, particularly in terms of the L2 competence attained and the risks involved in the development of the L1 (**Genesee1978**; **Genesee2013**; **Lazaruk2007**). Despite these concerns, the immersion education model has developed rapidly, inspiring bi- and multilingual school programmes throughout the world (**DeMejia2002**). The acknowledged success of immersion programmes may be due to a combination of factors that have been shown to positively affect L2 acquisition, such as onset age, the type of input made available and its quantity, that is, the amount of time allocated to L2 exposure, methodological flexibility (early, middle and late programmes) and teachers' backgrounds, among others (**Genesee2013**; **JohnsonSwain1997**; **Lazaruk2007**).

This chapter is part of a larger study Tejada-Sanchez **Tejada-Sanchez2014** which examines the outcomes of immersion programmes in Colombia, focusing on EFL writing of L1-Spanish speakers. More specifically, it seeks to understand the relationship between the allocation of time in the programme and the resulting learners' written performance in their target language, English. This relationship has not been sufficiently addressed in studies on school immersion contexts outside Canada, and even less so in the Colombian context. Earlier studies and compilations have underscored the importance of addressing the effect of the time factor, and more specifically intensive exposure experiences, within L2 instructional settings (**Muñoz2012**). Consequently, there remains a gap in the literature as to how language productive abilities benefit from such intensive instructional experiences. Undoubtedly, the number of uncontrollable variables within educational settings, such as individual differences, curriculum and context specifics, make this a particularly complex endeavour. For this study, data collection was conducted during class time in order to ensure the students' participation. It included written data and a background questionnaire which was used to control for individual variables such as age, L2 exposure and target language contact hours outside the school.

The chapter thus presents a descriptive study that evaluates the relative effect of different amounts of exposure to L2-English in two early partial immersion programmes in Colombia. We begin by reviewing the literature concerning time as an essential factor within immersion programmes, to then go on to discuss writing development in terms of the CAF triad, as well as the measurements adopted for profiling these dimensions. We then move on to present the methodology. Finally, results and analysis are outlined, followed by a discussion. Concluding remarks will focus on the implications of this study for L2 education and specifically curriculum allocation of languages within immersion programmes in non-English speaking contexts.

2 Literature review

2.1 Time as an intrinsic factor for immersion programmes

The question of the influence of the amount of target language exposure on language proficiency was raised quite early in the implementation of French immersion programmes in Canada. Carroll's contributions in the mid-sixties and seventies around the characteristics of immersion programmes were fundamental. Regarding the time-skill relationship, he asserted: "There are many factors which contribute directly to the effectiveness of French instructional programs (...) Organizationally, it is considered *that the key factor is the number of hours of instruction* in French (...) In other words, *the more hours a pupil spends in French, the higher level of achievement is likely to be*" (**Carroll1975**, cited in **Swain1981**: emphasis added). He identified a direct link between the volume of input made available to learners, quantified as time, and the overall L2 attainment. **Stern1985**, in turn, referred to a threshold regarding the number of hours likely to ensure a *bilingual* competence in an immersion context: at least 5,000 hours, but this account did not determine the characteristics of the learner involved in the programme, and did not make explicit the distribution of exposure time or its intensity, in terms of hours per week/month. Currently, the publications which explore time as a factor in the development of an L2 emphasize its importance but at the same time its intricate complexity. The conclusions that can be drawn from **Muñoz2012**'s (**Muñoz2012**) compilation demonstrate that, depending on how and where time is operationalized in language education, it can lead to a myriad of effects, from cognitive to socio-pragmatic, from global language features to discrete ones such as those addressed by the CAF dimensions. In this study, we focus on the parameter of accumulated time of exposure. This parameter refers to the global amount of time, in terms of number of hours, dedicated to L2 learning (**Stern1985**; **Genesee1978**; **Genesee2013**). It is usually required for the completion of a programme with a given target proficiency level, as defined for instance by the CEFR descriptors (**CouncilofEurope2001**). Regarding the accumulated time of exposure, immersion programmes are those where L2-contact time, along with content integrated instruction, is deemed essential for the programme's functioning (**CollinsEtAl1999**). Globally, immersion programmes have been traditionally described as beneficial for receptive skills (**DayShapson1988**), while their limitations regarding writing and accuracy have been frequently reported in previous research (**Lightbown2012**; **GermainEtAl2004**). In such respect, in written and oral expression, immersion learners often demonstrate a considerable influence of L1 grammar. Also, it has been repeatedly reported that learners would not start a conversation in the L2 spontaneously, unless when they are asked to do so (**Harley1992**; **Wesche2002**). Finally, it is suggested that even though productive skills appear to be distant from those of native speakers, learners in immersion programmes continue to make progress in the L2 (**Harley1992**; **Wesche1989**; **Housen2012**).

Particularly in terms of writing, the main topic in this study, contributions by Bournot-**Bournot-Trites2007**, **CollinsWhite2011**,**TurnbullEtAl1998** and **Lightbown2012** underscore learners' capability to communicate effectively but failing to reach native-like levels, for instance as regards lexical diversity and structural elaboration.

Summing up, it has been prevalently hypothesized that the more time students dedicate to learning the L2, the higher their level of proficiency will be (**Stern1985**), thus supporting the spread of instructed immersion and intensive programmes (**SerranoEtAl2011**; **Lightbown2012**). However, although the pioneer Canadian initiatives have been abundantly documented in the SLA literature, research is scarce as far as other countries are concerned. Hence, the current study seeks to shed light on the effects of EFL immersion in Colombia, by examining and comparing the effects on writing performance of students belonging to two programmes which differ in total number of hours and their distribution. Individual differences such as L2 exposure outside school, family bilingualism, and total amount of time in the school were also taken into consideration, but will not be discussed in this chapter.

2.2 L2 Written performance

Writing is a cognitively complex and multidimensional endeavour involving different stages and processes (**Manchón2013**; **Ortega2012**). In fact, this skill is understood as an 'interactive' process where various factors, such as genre awareness (stylistic organization and textual format) and mastery of content and language, are frequently activated and deactivated, according to writing pace and the needs of the composition process.

Creating a text comprises three main stages, namely, planning, formulation and revision (**Manchón2009**; **Manchón2013**; **SilvaMatsuda2005**). In the case of an L2, this activity is complicated by additional demands such as the search for the appropriate lexicon, grammar, discourse and other peculiar dimensions of the target language and culture (**Manchón2009**).

In this study, writing is seen as a genuine and meaningful way of communication in controlled L2 settings, such as the immersion school. Thus, in line with **Harklau2002**, **Ortega2012** and **Williams2012**, writing is a means of promoting permanent opportunities for practicing and revising L2 production in the classroom.

Two main approaches have been used to analyse writing in this study: quantitative measures for the three CAF dimensions and qualitative assessment using holistic ratings.

2.3 Complexity, accuracy, and fluency (CAF)

The quest for a developmental index to describe L2 performance has been a key issue in SLA research for decades now (among the first attempts, see e.g. **Larsen-Freeman1978**). Building on models of L2 proficiency (**Skehan2009**; **EllisBarkhuizen2005**, among others), Housen et al's 2012 volume elaborates on the potential of CAF as complementary dimensions of language performance and as a reliable approach to gauging L2 proficiency, as the three dimensions encompass the major areas of performance in an interlanguage system.

In this study, we adopt the CAF triad to assess writing performance in immersion contexts. Several contributions (**BultéHousen2014**; **HousenKuiken2009**; **HousenEtAl2012**; **Wolfe-QuinteroEtAl1998**), have discussed the operationalization of these measures in order to explain what makes a learner a *skilled* user of a language. Below we review those adopted for our study.

2.3.1 Complexity

Complexity is a construct that reflects the multidimensionality of the language learning process. It particularly poses numerous problems in the SLA field due to its polysemic nature, which can refer to structural, cognitive and developmental aspects (**Pallotti2015**).

In this study, L2 complexity is analysed from the language structure point of view claimed by **HousenEtAl2012** and **Pallotti2015**. This implies looking at the properties of L2 constructions, forms, form-meaning mappings and their interrelationships.

Several accounts have discussed the multiple operationalisations of this construct and underscored its problematic nature (i.e.**Wolfe-QuinteroEtAl1998**; **NorrisOrtega200** **Pallotti2015**; **HousenEtAl2012**; **Skehan2009**). In this respect, a wealth of measurements have been applied, revealing relatively operationalization vagueness and 'low content validity' (**BultéHousen2014**). Its multicompositional nature implies that complexity operates based on major assumptions that include: 'the more content means more complex', or 'the longer', 'the most embedded' or the 'more varied', all imply more complexity. As **BultéHousen2014** emphasize when examining short-term changes in written complexity, L2 research needs to be cautious about the validity of such measures and their implications, as their predictions may vary depending on the context, the learner and the task.

In light of these observations, this study seeks to adopt complexity as an indicator of L2 performance at different stages of language instruction. The selection of measures for syntactic and lexical complexity takes into account the nature of the texts produced by different groups of learners, which, in our study are often rather short.

2.3.1.1 Syntactic complexity

Syntactic complexity is generally measured through the length, proportion, combination and interrelation of different elements within a text (**BultéHousen2014**). Several elements or units have been taken into consideration such as the sentence, the clause and the T-Unit, among others. Following **Pallotti2015**, this study examines L2 syntactic complexity by analysing structural properties at the sentential and the clausal level, as well as text organisation properties through the use of coordination and subordination. Following **TorrasEtAl2006** the measurements adopted for the study were independent and dependent clauses per sentence (IndepCS and DepCS), and, following **BultéHousen2014**, the Coordinated Clause Ratio (CoordCR) calculated by dividing the number of coordinated clauses by the number of sentences was calculated. As argued by **BultéHousen2012**; **BultéHousen2014** this type of calculation (CoordCR) highlights the use of coordination within a text and differs from the Coordination Index developed by **Bardovi-Harlig1992** in that the CI appears to be a measure of clause combination that entails subordination as well: "the score on this index depends on the amount of subordination produced" (**BultéHousen2012**).

2.3.1.2 Lexical complexity

Lexical complexity has been frequently analysed by looking at lexical diversity, density and sophistication (**HousenEtAl2012**; **BultéHousen2014**). Diversity, also known as lexical range (**Crystal1982**) or lexical variation (**Read2000**), is measured through calculations which account for the variety of vocabulary items within a language sample (**MalvernEtAl2004**). Measurements of density and sophistication are mostly used either with larger text samples or to discriminate amongst text genres (**Read2000**). Nonetheless, it has also been argued that these measures do not really operationalize structural complexity. As **Pallotti2015** highlights, "indices of lexical sophistication, like the percentage of rare or difficult words, may be valid indicators of development, but they do not directly tap structural complexity; from a structural point of view, a rare word like *tar* is not in itself more complex than a common one like *car*." Today, there is a general consensus in that diversity, sophistication and density (and an additional dimension of lexical accuracy) allow us to profile vocabulary development. In addition, diversity has been frequently examined with shorter texts such as those in our data (**MearaMiralpeix2017**).This is then the measure we have adopted to assess lexical complexity in this study.

Thus, this study uses two measures of diversity to gauge learners' lexical repertoire. First, *Guiraud's Index,* which results from dividing the number of types by the square root of the tokens in order to limit text size effects. The second one is D, computed with the *vocd* tool in CLAN (**MacWhinney2000**), which estimates lexical distribution in longer text samples (**MalvernRichards2000**). Both measures have been used used to gauge language diversity in general; however, consensus has not been reached over which index proves to be a better predictor of lexical diversity in a person's interlanguage (**McCarthyJarvis2010**, in **Pallotti2015**). Therefore, this study will report both measures, to provide a more comprehensive picture.

2.3.2 Accuracy

The accuracy domain refers to the appropriateness of grammatical, lexical, semantic and pragmatic choices with respect to L2 target parameters. It is one of the most observed traits in the language production of L2 learners and it has been frequently treated as a key aspect of interlanguage development (**HousenEtAl2012**). Accuracy is operationalised by counting the grammatical and lexical errors in a linguistic production. However, **Polio1997**; **Polio2001** remarks that the most commonly used measure for this domain is the quantity of units with no errors

(Error-Free units), which poses problems for the analysis of short compositions or those by beginner learners. In this study, overall measures of specific errors such as total amount of errors per 100 words and grammar errors per 100 words (ToralErr/100 and GrErr/100) were calculated, as they capture the totality of errors produced as well as their structural category. Grammatical errors were predominant in most of the scripts, and they mainly corresponded to agreement phenomena and verb conjugations.

2.3.3 Fluency

This term is commonly associated with the speed of articulation, rhythm, and pausing in the production of oral language. In the case of written compositions, it refers to the length of units, that is, the quantity of words and structures produced within a given time (**BultéHousen2014**). To account for written fluency, this study adopts the view whereby the proportion of words produced is observed in relation to a given amount of time (task time, which in our case was 20 minutes). Previous research employed measures such as the number of units produced per minute, or the number of units produced within a 'macro' structure such as the sentence; in the present study, measurements in this domain include words per minute and words per sentence (WM and WS). These measures provide an account of fluency in terms of quantity and rate of production. These were chosen over analogous proposals such as *words per burst,* defined as the number of written words produced between two pauses or other interruptions (**Gunnarsson2012**), as the scripts analysed for this study were not collected using key-logging technologies.

2.4 Holistic ratings

Holistic approaches to the evaluation of L2 writing have frequently been used in SLA research. These can be operationalised through scoring carried out by trained raters following assessment rubrics. These instruments usually consist of descriptors of the language used by the learner as well as the degree of completion of a given task. For example, standardized tests' examination grids, (e.g. TOEFL) include various indicators that reflect a learner's L2 competence according to specific criteria, purpose and genre. In L2 research, these ratings often serve as complements to objective measurements of text quality (**Weigle2002**).

Our study uses a scoring rubric for the qualitative assessment of learners' written composition (**FriedlAuer2007**). This scale examines the characteristics of beginner to high intermediate levels of expository and narrative composition, also including task completion criteria. It was originally designed for the evaluation of English-L2 written performance within CLIL school settings in Austria (**FriedlAuer2007**; **Dalton-PufferEtAl2010**) and later on in Catalunya (**Juan-GarauSalazar-Nogue RoquetPérez-Vidal2015**). Four aspects are evaluated on a global scale of 0 to 20, which is in turn divided into four subscales ranging from 0 to 5: 1. Task fulfilment, 2. Text Organisation, 3. Grammar and 4. Vocabulary. In the current study this instrument has been adopted to profile learners' descriptive writing within content-based instruction contexts, which are highly comparable to the contexts for which it was originally developed (**Pérez-Vidal2013**).

3 Research question

This study explores the relationship between L2 exposure time and writing performance in immersion learners, as measured through the CAF constructs and holistic ratings. Hence, the guiding research question is:

1. Does accumulated time of EFL exposure in two contrasting immersion programmes (HI and HI+) have a differential impact in the long run on the learners' writing performance, when assessed with a) CAF measures and b) holistic ratings?

On the basis of this question and our review of the literature we hypothesize that, at any given time that learners are measured in the respective programmes, the higher the number of accumulated hours of EFL exposure students receive, the higher their level of proficiency will be.

4 Method

4.1 Context and participants

Foreign language education is a central theme in Colombia's political agenda (**BonillaTejada2016**). English plays a major role in a long-term education project entitled *Colombia Bilingüe*, which aims to rank Colombia as the highest provider of quality in education in Latin America.

Our study focussed on two immersion programmes with different total times of EFL exposure. We have named them High Intensity (N=52) and High Intensity *plus* (N=136) (HI and HI+) for the purpose of the study.[1] The difference in the

[1] The designation of these programmes has been adopted from **CollinsEtAl1999**; **Bournot-Trites2007**, and **CollinsWhite2011**.

number of participants in both programmes is due to a larger pool of students in the schools following the HI+ programme. Table 1 displays the number of participants per programme, age-group, and grade involved. Both programmes represent actual implementations of immersion models in the private sector of Colombia's educational system, with rather high amounts of L2 exposure compared to the average Colombian traditional EFL programmes. In the public sector, time of L2 instruction ranges in average from 2 to 4 hours per week, whereas in the private sector these amounts of L2-exposure are much higher, ranging from 7 to 20 hours per week, with L2 content-based instruction being predominant.

Table 1: Number of participants per programme.

.8SSSS

Age-group	Grade	N HI+	N HI
12	6th	12	14
14	7th	34	8
15	8th	22	8
16	9th	20	5
17	10th	48	17
Total	11th	136	52

HI and HI+ follow an early partial EFL immersion model in an otherwise Spanish curriculum, the official language in Colombia. Schooling begins at the age of four in kindergarten. From this age onwards, courses are taught about 50% of the time in the L2 and the other 50% in the L1. The most significant exposure to the L2 is offered mainly through immersion instruction, that is through curricular content taught in English. In neither programme is English taught through a grammatical or metalinguistic approach. Interestingly, students seldom use the L2 outside the classroom or for non-academic activities, so there is little or none informal learning.

Figure 1 displays the mean L2-instruction hours per year in both programmes. Black refers to HI+ and light grey refers to HI. Primary school, which lasts five years (1st to 5th grade), is the most intense period in terms of L2 exposure, as most of the subject areas (Sciences, History, Arts, etc.) are taught in English in both programmes. In terms of time distribution in primary school, HI+ provides between 600 and 670 hours of L2-exposure per year, whereas HI provides 504 hours. High-school (6th to 11th grade) is characterized by a decrease in L2-instruction time in both programmes. The HI+ programme offers 372 hours of

Figure 1: : Mean L2-instruction hours per school year for programmes HI and HI+.

L2-instruction per year by the end of this stage, while the HI offers 288 hours.

Regarding the curriculum at higher stages, both the HI+ and HI programmes coincide in that the only subject areas taught in English during high school are *English Language Arts* or *Anglo-Saxon Literature*. These are offered in the L2 from 9th grade on in both programmes (around age 15). Opportunities for exposure to English at other locations in the schools, for example in the school cafeteria, the playground or common areas remains limited.

The HI+ programme gathers students from three schools. These offer the largest number of hours of L2 exposure-instruction time: **8760 accumulated hours** by the end of grade 11 (age 17). At the end of a school in the HI+ programme, a renowned international certification is provided.[2]

The HI programme involves students from one school. It offers a relatively lower number of hours of L2 exposure-instruction time, **7002 accumulated hours** by the end of grade 11 (age 17). A singular academic approach to literacy in the HI curriculum is underscored by the school's stakeholders (principals, coordinators and teachers), so students are frequently exposed to discourse and text analysis since primary school. Table 2 shows the distribution of hours per year and the accumulated hours in the two programmes.

Table 2: Number of hours of L2 accumulated per year per programme.

l SSSSSS

Grade	**6th**	**7th**	**8th**	**9th**	**10th**	**11th**
Age	**12**	**13**	**14**	**15**	**16**	**17**
HI	**4914**	**5418**	**5922**	**6426**	**6714**	**7002**
HI+	**6312**	**6924**	**7536**	**8016**	**8388**	**8760**

4.2 Design and procedure

Written data was collected from five different age-groups (ages 12, 14, 15, 16, and 17) in each programme. Age 13 was not taken into consideration since the data collection process could not be completed with the whole group in one of the programmes.

[2] The *International Baccalaureate* certification (IB). In order to reach such a goal, these students from HI+ must follow the program for another year, grade 12, which was not considered in this study.

4.2.1 Data collection and trimming

Two main instruments were used to collect the data used for the present study: a linguistic background questionnaire and a written task consisting of a composition based on a silent film.

4.2.1.1 Linguistic background questionnaire

A general linguistic background questionnaire inspired by **Grosjean2010** was used to investigate participants' use of different languages, their learning habits, their L2 interaction and contact with target language speakers, as well as the average time spent in the immersion programme. This instrument was later used to make a selection of participants in the study. Students who had not been in the same school for their complete tuition (from primary years onwards), had lived in an English-speaking country or abroad, were binational or had English-speaking relatives, were excluded from the study. This left a final sample of 188 students including both programmes, as shown in Table 1.

4.2.1.2 The writing task: Retelling a story

In order to collect data on the participants' written abilities, they were asked to write a story retell on the basis of the silent film "College" (**Keaton1927**), starring Buster Keaton. The choice for this task emerged from earlier analyses on task structure such as **SkehanFoster1999**, where narrative retellings tasks supported by visual prompts are used to elicit the three dimensions of CAF in comparable degrees. Likewise, silent films have frequently been used in SLA studies to elicit narratives in the L2 (e.g. **Lambert1997**, who used Chaplin's *Modern Times*).

Participants watched a 3.30-minute scene of the film, which was played only once. Subsequently, they were allowed 20 minutes to complete the composition. They were asked to write as much as they were able to in the given time. They received the following instruction:

is this numbering intended

1. Retell the story in writing while keeping in mind all the details. Use your current knowledge of English; do not use the dictionary.

4.2.2 Data coding and analysis

All the participants' compositions (N=188) collected through the written tasks were transcribed and coded using CLAN (**MacWhinney2000**). A first streamlining was conducted to standardize coding procedures. L2 errors and spelling occurrences were identified and scripts were segmented into units. The errors that were not taken into account were those caused by phonology or graphical ambiguity (i.e. *the man say's*), misspelling (i.e. *he whent*), redundancy, or repetition of text content (i.e. ***A man* put a poster that says Boy needed. And then *a man* come and tell that he want the work**).

CAF measures and holistic ratings were employed to analyse the learners' writings. CAF analysis was carried out through manual coding of grammatical and lexical errors, segmentation of syntactic units and automatic calculations using CLAN and Excel. Holistic ratings were carried out by two external evaluators. In order to compare learners' performance in terms of the impact of the accumulated time of exposure in HI and HI+, descriptive statistics (means and SDs) were calculated on both CAF measures and holistic ratings for all the age groups combined (12, 14, 15, 16, and 17) in each programme, which allowed us to measure the effects throughout the programme. Between-groups comparisons were conducted using Welch's t-test.

4.2.2.1 The unit of analysis

The main unit of analysis in our study is the sentence. Our scripts resulted in an average of 32 words (see Table 3), which made an analysis based on T-Units (**Hunt1965**) too restraining, as this syntactic unit requires longer compositions to allow for a more substantive examination of how the units are conceived by the writer in terms of length and interrelations between clauses. Following **Bardovi-Harlig1992**, this study adopts the sentence as the main syntactic unit in order to keep the author's original textual/syntactic segmentation.

We followed the criteria for defining the sentence and the clause established by **GreenbaumQuirk1990**,**Bardovi-Harlig1992**, **Vyatkina2012** and **BultéHousen2014**. We understand the sentence ('S') as a 'grammatically autonomous unit' (**QuirkEtAl1985**) having a subject, at least one conjugated verb and possible complements. In written texts, sentences are identified as those stretches of writing enclosed between two full stops, or between a full stop and a colon or semi-colon. The clauses ('C') are the units which combined together form different types of sentences: simple, compound and complex. They contain a subject and predicate and can be independent or dependent (subordinate). Likewise, according to **BultéHousen2014** "a sentence can also include two or more coordinated independent clauses and become longer by adding more coordinated and/or subordinated clauses, when their constituent clause(s) contain more constituents and phrases, and when the phrases that make up these clauses contain more words" (p. 49). In contrast, a T-unit consists of one independent clause with all of its dependent (subordinate) clauses and they do not become longer when coordinated clauses are added.

The following excerpts are sentences derived from the data examined, and they serve to illustrate the segmentations applied for this study. T-Unit boundaries have also been marked (/). Sentence length differs between both subjects as does the amount of coordinated (Coord) and dependent clauses (DepC). L2-errors have been kept as in the original.

Excerpt 1 (Grade 10, Age 16, HI): Ronald passed throw [= through] the store and saw an announcement that says Boy Wanted, / so he decided to enter in the store and ask for the job.

(1 S, 2 T-Units, 2 Coord, 1 DepC).

When Ronald saw a beautiful girl in a table he was ashame of working as a clerk / so he went out of the bar and sat down as if he was a client.

(1 S, 2 T-Units, 2 Coord, 1 DepC).

Excerpt 2 (Grade 6, Age 12, HI+): There was a man that get by train to a new city.

(1 S, 1 T-Unit, 1 DepC).

he don't have the good cordination to do it /so he say* that he cannot do it again and he go*.

(1 S, 2 T-Units, 2 Coord, 1 DepC).

Based on this analysis, the scripts examined for this study were fairly short, as shown in Table 3, with an average of 32 words, 5 sentences, and 11 clauses.

Table 3: Main descriptive statistics for the whole corpus

Xrrrr

Words	**Sentences**	**Coordination**	**Dependent clauses**	**Independent clauses**
31.54	**4.98**	**3.23**	**4.29**	**10.51**

4.2.2.2 CAF measures

A total of nine measures, in the form of frequencies, means, and ratios, were examined in this study to account for complexity (syntactic and lexical), accuracy (use of L2 target parameters) and fluency (quantity of words). Table 4 presents the summary of the measures adopted.

Table 4: Summary of CAF measures applied in this study

llQ

Domain	**Subdomain**	**Measures**
Complexity	**Syntactic**	**Independent clauses per sentence (IndepCS)**
		Coordinated Clause Ratio (CoordCRatio)
		Dependent clauses (DepC)
	Lexical	**Guiraud's Index**
		D
Accuracy		**Errors per 100 Words**
		Grammar errors per 100 words (GrErr/100)
Fluency		**Words per minute (W/M)**
		Words per sentence (W/S)

4.2.2.3 Holistic ratings

About 55% percent of the scripts (100 in total, 10 per age-group and 50 per programme) was assessed by two evaluators from different backgrounds (Table 5). Rater 1 was a female EFL teacher in Colombia, she is originally from Cincinnati, Ohio (L1-English and L2-Spanish). Rater 2 was a female EFL teacher from Colombia (L1-Spanish and L2-English). Each evaluator scored all the narratives according to the chosen scale without knowing the authors' age or programme. Inter-rater reliability was examined by calculating the intra-class correlation (ICC) for the two programmes in each criterion of the rubric. Evaluators' agreement in scoring each immersion programme was moderate to strong on most of the rubric's criteria, except for Text Organisation. In this case, the ICC obtained for HI+ was 0.33 and for HI it was 0.66.

5 Results

5.1 CAF measures

The outcomes of CAF analysis for both programmes are shown in Table 6. Three measures are used for syntactic complexity: independent and dependent clauses per sentence, and the Coordinated Clause Ratio (IndepCS, DepCS and CoordCR). In terms of all these measures, the HI+ group has lower figures than the HI group, which appears to produce slightly more coordinations and subordinations throughout its scripts. Lexical complexity, as measured by D and the Guiraud index, proves to be similar in both groups, with relatively low values of D (between 42 and 43). As regards accuracy, the Errors per 100 words (Err/100) measure shows similar results for both programmes. Regarding grammar errors per 100 words (GrErrr/100), HI+ students produce an average of 9.09 errors per 100 words and the HI students produce 11.72. Lastly, fluency measured through the number of words per sentence (WS) appears higher in the HI group, while it his slightly higher in the HI+ group when measured in terms of words per minute (WM) (in a 20-minute task).

Welch's t-tests were conducted to assess the statistical significance of between-group differences. Table 7 reports on the results of these tests as well as the effect sizes through Cohen's *d*, which were small to medium, according to **PlonskyOswald2014**'s (**PlonskyOswald2014**) suggested criteria. In terms of Complexity, HI+ and HI prove to be significantly different as far as the production of independent clauses (IndepCS) (t=3.700, p < .001), with the HI + group producing fewer independent clauses than HI. Likewise, groups appeared to be significantly different concern-

Table 5: Holistic ratings and Intra-class correlation for both evaluators and programmes

>p18mm SSr SSr

	HI+ (n=50)			HI (n=50)		
	Rater 1	**Rater 2**		**Rater 1**	**Rater 2**	
	Mean *(SD)*	**Mean *(SD)***	**ICCHI+**	**Mean *(SD)***	**Mean *(SD)***	**ICCHI**
Task fulfillment	2.47 (0.80)	3.32 (0.81)	0.54	2.64 (1.14)	3.20 (0.87)	0.72
Text organisation	2.00 (0.62)	2.71 (0.69)	0.33	2.24 (1.02)	2.78 (0.76)	0.66
Grammar	2.57 (0.69)	2.66 (0.75)	0.58	2.62 (0.86)	2.64 (0.77)	0.78
Vocabulary	2.48 (0.73)	2.58 (0.76)	0.66	2.53 (0.92)	2.47 (0.79)	0.68
Total score	9.52 (2.32)	11.27 (2.76)	0.60	10.01 (3.44)	10.97 (2.91)	0.81

Table 6: Descriptive statistics for all CAF measures for programmes HI+ and HI

l@Q r@ r@ r@ r @ r@ r@ r@ r

Measure	**mean**	***sd***	**Min.**	**Max.**	**mean**	***sd***	**Min.**	**Max.**
IndepCS	2.68	0.76	1.2	7	3.26	1.03	1.86	7
CoordCR	0.59	0.37	0	10	0.68	0.53	1	10
DepCS	0.56	0.31	0	1.5	0.75	0.44	0.15	2.25
Guiraud	1.52	0.27	0.76	2.35	1.46	0.19	1.06	1.8
D	42.69	13.21	19.33	98.21	41.78	8.65	26.08	72.42
TotalErr	15.85	8.16	0	33.33	17.05	7.39	0	30
GRErrors100	9.09	5.48	0	20.83	11.72	6.80	0	26.78
WS	6.20	1.81	3.286	16	7.25	2.37	4	15
WM	2.24	0.81	0.47	4.07	2.07	0.73	0	0.41

ing subordination (DepCS), where HI+ pupils appears again to write fewer dependent clauses than HI (t = 2.868 p < .05). Regarding both measures of lexical complexity (D and Guiraud index) no statistical differences were found.

Concerning accuracy, the calculation of grammar errors per 100 words (GrErr/100) yields significant differences between groups. The HI+ subjects seem to produce significantly fewer errors than their HI counterparts (t = 2.494, p < .05).

Finally, as per fluency, HI and HI+ learners significantly differ in terms of the words produced per sentence (WS), where the HI+ programme used around one word less per sentence when compared to HI (t = 2.887, p < .05). No significant differences were found as regards words per minute.

These results could be summarized by noting that HI+ learners produce fewer independent and dependent clauses, fewer words per sentence, but fewer grammar errors per 100 words than HI. That is, they are less complex and fluent, but more accurate. These findings could imply a trade-off effect. In terms of lexical complexity, both groups appear to perform similarly.

5.2 Holistic ratings

Figure 2 shows two graphics with the two evaluators' scores for programmes HI and HI+ on the Total Score of the rubric. Rater 2's scores appear to be systematically higher than rater 1's. These discrepancies might be attributed to 1)

Table 7: Results for Welch's T-Test for between-group comparison of programmes HI and HI+

Ql Srrr

Domain	**Measure**	**Statistical value (/t)**	***p***	**95% CI**	***d***
Syntactic Complexiy	**IndepCS**	**3.700**	**p < .001**	**0.267 0.890**	**0.60**
	CoordCR	**1.119**	**0.26**	**-0.070 0.249**	**0.18**
	DepCS	**2.868**	**p < .05**	**0.058 0.324**	**0.46**
Lexical Complexiy	**Guiraud**	**-1.834**	**0.068**	**-0.135 0.005**	**-0.3**
	D	**-0.553**	**0.58**	**-4.175 2.348**	**-0.09**
Accuracy	**TotalErr/100**	**0.969**	**0.34**	**-1.258 3.663**	**0.15**
	GrErr/100	**2.494**	**p < .05**	**0.530 4.726**	**0.40**
Fluency	**WS**	**2.887**	**p < .05**	**0.325 1.772**	**0.47**
	WM	**-1.387**	**0.16**	**-0.414 0.073**	**-0.22**

the evaluators' different L1 backgrounds and 2) a differential judgement of text structure, grammar and lexical repertoires (raters might have judged learners' lexicons not only in terms of diversity but in terms of accuracy).[3]

Interestingly, the scores don't seem to change much across different age groups, except for a slight positive difference between initial (age 12) and final (age 17) levels. Both raters judged scripts produced at age 16 with the highest scores, with a rather surprising decrease at age 17.

Figure 2: Rater 1 and Rater 2 Total Scores attributed to the scripts from HI and HI+ based on a 20-point scale rubric

Between-group comparisons using Welch's t-test did not reveal any statistical differences between the programmes, as shown in Table 8. The mean difference between raters' perception of HI+ and HI on various aspects of writing ability ranges from -0.17 to 0.03. These results suggest that neither programme is perceived as significantly different from the other, when it comes to the holistic

[3] Open-ended questionnaires have been used in SLA research in order to explore raters' assumptions and beliefs (see for example by DelRioEtAl2018tv), which could be a possibility for further research on our corpus.

rating of L2 writing performance.

Table 8: : Analysis of between-group differences in holistic ratings by two evaluators (Welch's T-Test)

Q rrr rrrr

	[2cm]Group HI+ Mean (SD)	**Group HI Mean (SD)**	**Mean difference between groups**	***t***	***p***	**95% CI**		***d***
Task fulfillment	**2.895 (0.69)**	**2.928 (0.93)**	**-0.033**	**0.191**	**0.84**	**-0.293**	**0.355**	**0.03**
Text organisation	**2.355 (0.51)**	**2.525 (0.82)**	**-0.170**	**1.15**	**0.253**	**-0.112**	**0.422**	**0.23**
Grammar	**2.615 (0.60)**	**2.628 (0.74)**	**-0.012**	**0.008**	**0.992**	**-0.264**	**0.266**	**0.001**
Vocabulary	**2.530 (0.64)**	**2.50 (0.74)**	**0.030**	**-0.334**	**0.738**	**-0.317**	**0.225**	**-0.06**
Total score	**10.395 (2.17)**	**10.520 (2.97)**	**-0.125**	**0.161**	**0.871**	**-0.939**	**1.106**	**0.03**

6 Discussion and conclusions

This study has sought to understand whether the differential accumulated time of EFL-exposure (expressed in number of hours of L2 learning) has an impact on writing performance in two immersion programmes, HI+ and HI. They are different in the accumulated number of hours at all points throughout the programme, and clearly at the end, at learners' 17 years of age, when the HI+ programme has accumulated 8,760 hours, while the HI programme 7,002.

CAF measures and holistic ratings of writing samples were scrutinised with a cross-sectional design in which learners were measured throughout the programme, on a yearly basis, starting at age 12. Concerning CAF, four measures out of nine (IndepCS, CoordCR, DepCS, Guiraud, D, TotalErr/100, GrErr/100, WS, WM) were found to be statistically different between programmes but not all in favour of HI+. As regards complexity, IndepCS and DepCS were significantly lower for the HI+ group; for accuracy, GrErr/100 were statistically higher for the HI+ group; for fluency, WS, again, was statistically lower for the HI+ group. In terms of lexical complexity and the holistic ratings, no significant differences were found between the two programmes.

Overall, it would seem to be the case that the two programmes are not substantially different in terms of learners' outcomes in EFL written performance. However, it cannot be said that they are entirely the same either. Indeed, the HI+ programme reveals lower levels in the domains of syntactic complexity and fluency, but higher levels for accuracy, and equal levels for lexical complexity. This has been also found in studies on the effects of a CLIL course in English added to conventional formal instruction contrasted with a group only taking formal instruction, as the latter outperformed the former, although not significantly (**RoquetPérez-Vidal2015**).

Consequently, given its mixed results, this study partly questions the early assertions made by **Carroll1962** and **Stern1985** in the direction that more L2-exposure time would directly lead to more skilled language use. Our approach to the interpretation of these findings is in terms of time distribution of each of the two programmes, between learners' ages 12 and 17, as presented in Table 2 and Figure 1. In the case of HI+, learners undergo a decrease of L2-exposure time which goes from 672 hours a year to 612, and then to 480 (see Figure 1). This is not the case for HI pupils, who receive fewer hours of target language exposure per year, 504, yet at a steady rhythm. Additionally, the reduction in exposure is placed one year earlier for the HI+ group, that is at age 15, than for the HI group, at age 16.

On the one hand, such a contrast in the distribution of L2 exposure time in the two programmes allows us to suggest that gradual exposure to the L2 (HI programme) might explain the similarity in results with HI+, with more accumulated amount of L2 exposure yet less consistent in its distribution.

However, such a constant exposure experienced by the HI learners may also have had a less positive consequence; that is, the HI learners' relative lower scores in terms of grammatical accuracy. In this sense, the notion of *stabilisation*, or *plateauing*, proposed by **Long2008** might be relevant. Indeed, a closer analysis of the learners' performance suggests a plateau effect mainly concerning grammatical accuracy, in the case of the HI programme predominantly observed in conjugation and agreement errors, a finding which has already been identified in immersion learners in the literature (**Rifkin2005**; **HartEtAl1991**). HI's outcomes in accuracy could be interpreted as a level of "maintenance" achieved in this programme. These findings can be relative to the regular and steady amount of exposure for HI students in primary and between ages 12 and 15 in secondary school, as exposed in Figure 1.

On this note, **Bournot-Trites2007**'s (**Bournot-Trites2007**) findings are only partially confirmed in our case. In her study, no significant differences in writing quality were found between two groups of secondary immersion students with different L2-French intensity. In addition, **Bournot-Trites2007**'s (**Bournot-Trites2007**) study reveals a plateau effect in the field of grammar accuracy (particularly tense markers) and lexical diversity, where she observes: "it seems that after a certain threshold of competence in [L2], the increase in the time spent in this language in the class does not improve much the quality of the written production of pupils" (**Bournot-Trites2007**).

Likewise, the similarity of the two programmes in terms of lexical complexity could also be explained in terms of input exposure. It would seem that neither programme offers complementary hours of exposure outside of the classroom which would aid learners to make progress in such a domain.

Concerning the lower levels of fluency found in the HI+ group, they could be attributed either to the programme's didactic approach, or to task effects which remain to be explored in future research.

Some limitations of this study need to be acknowledged. First the unbalance in the sample size where the HI+ programme includes a larger number of subjects than HI, which is represented by fewer subjects. Second, task conditions as well as task variety (in terms of complexity and text genres) need to be reconsidered. Further research might include different types of tasks and writing genres with different cognitive demands. We should additionally underscore that the HI programme-related positive results, which refer to denser, richer texts, may be associated to the emphasis on literacy in the HI's curriculum described in section 4.1.1. The task might have been more familiar to HI students and therefore yielded to slightly more syntactically complex and organised texts.

To conclude, the present study has confirmed that the examination of the time factor in L2-acquisition in formal educational settings remains a rather complex endeavour due to a number of methodological constraints and issues. It is difficult to assess different programmes at exactly the same times (in terms of age, L2 exposure, curriculum years), and to control for programme features. Future research is needed to pursue research in bilingual schools or immersion programmes in non-English speaking contexts and to explore performance differences among different age groups, with a mixed methods approach including the holistic analyses suggested.

Acknowledgments

The authors are grateful to the anonymous reviewer as well as to Gabriele Pallotti for their insightful comments and contributions to this chapter. Also we would like to thank Cylcia Bolibaugh for her statistical advice. Isabel would like to thank Alejandra Plata and Alejandro Mejia from Uniandes for their help with earlier versions of this text.