A Study of Learner Profiles in Spanish as a Second Language in a Swedish Instructional Setting: Writing versus Speaking

Berit Aronsson (Umeå, Sweden)

Abstract (English)

The present study evaluates test results of oral and written production for a group of 15-16 year-old Swedish L2-learners of Spanish in an instructional setting. The test results are evaluated according to the Swedish grade descriptors for the tested level and the descriptors of the Common European Framework of Reference for Languages (CEFR). It is shown that the subjects generally perform better in writing than in speaking. This result is valid both with respect to the Swedish descriptors and the CEFR scale ratings. The results are discussed in relation to the conditions for learning an L2 in a formal instructional setting with little exposure to the target language outside the school context. A possible dominance of writing-based activities in the classroom and a low degree of extramural exposure are suggested as possible explanatory factors to the results.
Keywords: Adolescent L2-learners, oral and written production, instructional setting, language testing, CEFR

Abstract (Español)

El estudio evalúa el resultado de una tarea de producción oral y escrita del Español Lengua Extranjera (ELE) realizado por aprendientes suecos de 15-16 años en un contexto de aprendizaje formal. Los resultados se evalúan de acuerdo con los descriptores de calificación suecos para el nivel evaluado y los descriptores del Marco Común Europeo de Referencia para las Lenguas (MCER). Los resultados demuestran que los sujetos generalmente se desempeñan mejor en la escritura que en el habla. Este resultado persiste también en la comparación con las clasificaciones de la escala MCER. Los resultados se discuten en relación con las condiciones para aprender una L2 en un entorno de instrucción formal con poca exposición a la lengua meta fuera del contexto escolar. Se sugieren un posible dominio de actividades basadas en la escritura en el aula y un bajo grado de exposición extramural como posibles factores explicativos de los resultados.
Palabras clave: Aprendizaje L2, producción oral y escrita, contexto de aprendizaje formal, evaluación de lenguas, CEFR

1 Introduction

The fact that linguistic proficiency is composed of several subskills becomes especially salient when we learn a foreign language. Learner profiles are often *spiky* (Ericksson & Börjesson 2001: 258, Smith 2016: 190). A spiky learning profile means that a learner has divergent levels of skills in different subdomains of an overall area. The present study evaluates and compares the linguistic level achieved in oral and written production in a group of Swedish L2 learners of Spanish after finishing school year 9 (Secondary School). The tasks tested have been evaluated according to the Swedish grade descriptors for the tested level and the descriptors of the Common European Framework of Reference for Languages (CEFR). The results are discussed in relation to the conditions for learning an L2 in a formal instructional setting with little exposure to the target language outside the school setting.

2 Perspectives on the Teaching and Assessments of L2 Akills

2.1 Learner Profiles in Receptive and Productive Skills

L2 learners' proficiency is frequently unbalanced: a good receptive skill does not imply that a learner is good at writing, and being good at writing does not automatically imply that he or she is good at speaking. As a token of comparison, it is common that learners' receptive skills are more developed than their productive skills (e.g. Erickson & Börjesson 2001 and Ginther & Yan 2018).[1]

Various studies show that learner performance is incongruent as well in the productive skills of speaking and writing (e.g. Weissberg 2000, 2006, Pérez Vidal et al. 2012, Smith 2016). However, when it comes to holistic comparisons of students' productive abilities, investigations still seem to be scarce. Researchers tend to have their expertise in either one of the fields, and their research therefore often focuses upon a single language skill, rather than the whole picture. Cross-modality research is found, for example, in emergent literacy studies, focusing on what we can learn about L2 writing pedagogy by looking at the oral discourse that surrounds L2 writing activities (e.g. Kroll & Vann 1981, Weissberg 2006), whereas studies that compare L2 learners' skills in speaking and writing with the aim of investigating learner profiles appear to be few in number. Some diverging results from studies with both quantitative (*language testing*) and qualitative (*profile development*) approaches are presented below, among which Smith (2016) is the only one that includes the same age group as the present investigation.

In a North American setting, Weissberg (2000), in a small-scale study of five adult L2-learners of English with Spanish as the L1, investigated an aspect of the developmental linkage between L2 writing and L2 speech, namely morpho-syntactic development. The results pointed to a clear preference for writing over speech. Baba et al. (2013) study showed similar results for a group of 26 Japanese learners of

[1] See also the European Survey on Language Competences (ESLC 2012), where more subjects reached the highest CEFR level included in the study, i.e. B2, in listening and reading than in writing.

English. The great majority (70%) of the subjects said that they felt better when writing than when speaking and that they indeed preferred writing. It is noteworthy that among the remaining 30% who reported to prefer speaking, all but one of the subjects had experiences from studies abroad, which, thus, seems to have been a decisive factor (Section 2.3). The overall preference for the written modality in these studies may originate from the traditional view of learning in educational settings, in which writing-based activities have been central to teaching. The balance between writing and speaking activities in the L2-classroom, i.e. in settings in which extramural (out-of-class) exposure is low, will be addressed as a possible factor of influence in the shaping of learner profiles. This view will be further developed in the following sections.

Ginther & Yan (2018) studied the relationship between language proficiency, in terms of receptive and productive skills, and academic success in a North American context. The study followed a quantitative approach and included 1900 Tests of English as a Foreign Language (TOEFL) taken by Chinese students prior to entering the University of Perdue (Indiana, U.S.). The subjects' TOEFL results were generally higher for receptive skills (listening and reading) than for productive skills (writing and speaking). An additional finding, which is not further discussed by the authors, is that the test-results also consistently show a slightly higher outcome for writing than for speaking (Ginther & Yan 2018: 285), although these differences are not as salient as those between productive and receptive skills. The subjects included were adult international students who (presumably) learnt the foreign language in an Asian / Chinese learning setting. Since we do not know anything about the subjects' learner backgrounds, their length of residence in the USA and their previous exposure to the target language, it is close to impossible to speculate about underlying explanations to the outcome. As pointed out by the authors, many factors may be involved here, such as students' study habits, motivation, persistence, and integration into the larger academic and social communities. Not least, factors such as students' degree of extramural exposure in relation to the number of hours of formal instruction in the L2 and the quality of this instruction seem to be at play.

In contradistinction to the above-mentioned findings concerning adult L2-learners in an American learning context, Smith (2016) investigated the same age group as that of the present study, in a European setting. She analysed written and oral exams of nearly 40,000 Greek learners of L2 English and found that an overwhelming majority of test-takers scored significantly higher in speaking than in writing. Learners were tested at the B1, B2 and C1 CEFR levels, through which these differences persisted. The fact that English has an established role as a *lingua franca* in Greece and is present in speakers' everyday life might be a contributing factor to the results. Films are subtitled and not dubbed, and English is widely used as a signal of status and prestige. Moreover, more than eight million tourists visit Greece every summer, over one million of whom are British, and the vast majority of these tourists communicate with their Greek hosts in English (Oikonomidis 2003, Dimitrakopoulou 2017). In Greece, as in Sweden, it can therefore be assumed that the exposure to extramural oral English is relatively high, and Greeks may be even more directly in contact with spoken English through tourism than Swedes are. So, if we consider the holistic context in which English is learnt in Greece, the strong results for speaking are not surprising.

2.2 Foreign Language Testing in Europe – especially with Respect to Oral Skills

In the European Survey on Language Competences (ESLC, 2011), carried out by the European Commission with the aim of providing comparative data on foreign language competences in the participating European countries, Spanish as an L2 for Swedish learners was, for the first time, included in a pan-European comparison. While Sweden performed at the top in L2 English, the results for L2 Spanish were found at the bottom of the list. The level of English among Swedish adolescent learners has, for a long time, been high in international comparisons. In the European report on the effectiveness of English teaching (Bonnet 1998), for example, Sweden was found at the top, and in a follow-up report from 2004, mainly built on the same tasks (Bonnet 2004), Sweden, together with Norway, achieved the highest results. However, although spoken interaction is emphasized as important both in the Swedish curriculum and in the CEFR, oral performance was not tested in any of these reports.

The lack of productive and interactive oral tasks was considered a lack and a validity problem in relation to the national curriculums according to Bonnet (2004), as well as in a follow-up discussion by the Swedish Board of Education (Skolverket 2004). Despite this criticism, spoken production was also excluded from the ESLC (2011). Along the same line, an overview over national tests in Europe (European Commission 2015) reported that speaking is the least tested skill, while the most commonly assessed skill is reading, closely followed by writing and listening which are both tested to a similar extent (European Commission 2015). The alarmingly low results for Spanish in ESLC (2011) have been analysed by the Swedish Board of Education (Skolverket 2012, 2013) and by Bardel, Ericksson & Österberg (2019) for example. Various explanations to the poor result for Spanish as compared with English have been suggested; among others the shortage of certified teachers of Spanish and the low degree of students' exposure to Spanish outside the school setting in comparison with English.

2.3 Exposure to the Target Language Inside and Outside the Classroom

The degree of exposure (both input and output) to the target language has commonly been mentioned as an important factor for L2 language development (Muñoz 2012 for an overview). The degree of input to the target language in the instructional setting is indeed very scarce in comparison to a naturalistic setting with full immersion. If we compare the exposure to the first language (L1) to that of an L2 learnt only in the school context, the difference is, of course, overwhelming. Clark (2003: 41) calculates that the potential exposure that children have to their L1 could be estimated at around 70 hours a week, or 3,650 hours a year, while the exposure to the target language in the foreign language classroom typically range around three to four hours a week.

In formal L2-learning, not only the degree but also the type of exposure to the second language in- and outside the classroom setting seem to be crucial. Muñoz (2014) studied the influence of onset age and input amount on the performance of second language learners in instructed settings. Interestingly, learners' progression was measured in terms of their oral performance. The author found that that the best predictors

of L2-learners' oral performance were hours of immersion from abroad experiences and current informal contacts with the target language outside the school environment, while input alone, in terms of the number of hours of instruction turned out to be rather a weak predictor of L2 performance (Muñoz 2014: 473-474). Even though the author does not point this out, a possible explanation to this outcome might be the way the teaching is carried out. If the exposure to the target language is limited to the school context, and if the exercise types carried out in this context are largely based on writing activities at the cost of oral practice (as discussed below), the number of hours of instruction would be expected to be less influential on oral L2 performance.

As already suggested, one of the main differences between Spanish and English as foreign languages in Sweden is the degree of exposure to the language outside the school setting. While the exposure to English through different media is indeed high outside school, Spanish input is limited. To a great extent, English as an L2 in the Swedish setting is learnt not only in the classroom, but also through extramural activities, for example through television and computer gaming (e.g. Sylvén 2004, Sundqvist 2009, Olsson 2011, Sundqvist & Sylvén 2012, Sylvén & Sundqvist 2012). Olsson (2011) found a significant correlation between the frequency of extramural English and grades in English: the more frequent extramural exposure, the higher the grades. Sylvén & Sundqvist (2012) found that English interaction proficiency correlates with the frequency of gaming and the types of games played. Sundqvist (2009) found that digital gaming along with using the Internet, which could be seen as both productive and interactive activities, enhanced L2 learning more than the practising of receptive skills, such as, for example, listening to music or watching TV or films. This input is comparable to the input conditions mentioned as successful by, for example, Muñoz (2014), i.e. current informal contacts with speakers of the target language and immersion from abroad experiences, in the sense that all these activities require oral interaction with a counterpart. English is present in Swedish everyday life, television, songs, gaming, instruction manuals, computer instructions, etc., while this is not the case for Spanish. Since input from Spanish outside the classroom is scarce in comparison with English, pupils are probably limited to a great extent to the output / input that is provided to them through teaching, learning materials and the classroom environment.

2.4 The Dominance of Writing-Based Activities in L2-Teaching Materials and Assessments

As pointed out by Pérez Vidal et al., formal instruction typically revolves around writing and / or receptive, rather than productive oral skills (Pérez Vidal et al. 2012: 215). Due to the large impact of literacy and the predominance of written language regarding our way of thinking and reflecting on language (e.g Harris 2001 and Linell 2005), we may not even notice that many oral exercise types also involve reading skills (Aronsson 2014, 2016). The conception of language as being closely connected to writing and literacy may also be one of the explanations to the Cinderella-like treatment of spoken language in language testing. An overview of national tests in Europe (EACEA 2015) points out that reading and writing, as well as listening skills, are covered by the written parts of national language testing, since the test-takers make their answers in the written form. I argue that also the speaking test, at least the way it is carried out in Sweden, is to some extent based on a combination of writing and speaking. The

instructions of the tests are presented to test-takers in writing, and test-takers are encouraged to annotate keywords before starting the speaking session. Pictures, speech balloons and cards where different topics to discuss appear in writing are used to encourage speaking. In this sense, the test-format is not a true test of the test-takers' abilities in oral interaction, but rather of learners' conjoint understanding of the written and spoken form of language. Sandlund & Sundqvist (2011), who analysed turn-taking in an oral national test of English, reported that test-takers were affected by the written elements of the test-format. Thus, there were interruptions in the dialogues when test-takers asked the teacher questions about the topic cards (*Should I take the* next *[topic card]*), and when the teacher interrupted the test-takers' dialogue with comments (*Take another card then*), whereupon the test-taker responds with a question *e.g. Is it my turn*?) (Sandlund & Sundqvist 2011: 100-105). It cannot be excluded that learners with reading / writing disorders are discriminated against by this test-format, even though the test intends to measure oral proficiency.

The dominance of writing-based activities in assessments is by no means a new phenomenon in the field of language testing (Isaacs 2016 for an overview). While writing has been one of the traditional skills to be tested in language assessments, speaking has been included only more recently, initially in the form of optional supplementary tests (Spolsky 1995, Ericksson & Börjesson 2001, Isaacs 2016). L2-speaking tests were initially based on the translation of sentences or texts that were read aloud, as reported by Kaulfers (1944) in his study of assessments of foreign languages in wartimes. The author admitted that the proposed test format may expose the shortcomings of "silent pen-and-ink exercises", and therefore assess speech only indirectly. Read-aloud tasks were also used later on in order to avoid content as a source of variation and to access the linguistic properties of speech only (Isaacs 2016). The fact that international European investigations of L2 performance have at various occasions given priority to the testing of writing skills while excluding learners' oral performance give important signals to the practical field of language teaching: what matters is to be good at skills based on, or connected to, literacy. Bonnet (2004) expressed concerns regarding this matter as follows:

> Oral production seems to be regarded as an important skill in the curricula of all the participating countries. If we want to make sure that it is effectively trained and developed in teaching, testing of speaking should also be included in evaluations because of the expected wash-back effect (Bonnet 2004: 124).

Against this background, L2 learners in formal educational contexts with limited access to the target language outside the classroom may run the risk of primarily being exposed to the written form of language. Hypothetically, this would lead them to acquire writing skills at the cost of their oral proficiency – a development opposite to that of the mother tongue, where oral language development precedes the acquisition of writing. Oral skills are likely to develop before writing abilities even in a second language learnt in a naturalistic setting or with high extramural exposure to the target language, as is the case for the learning of English in a Swedish setting. If the classroom activities that Swedish L2 learners of Spanish participate in are, to a great extent, based on writing activities, our subjects' writing skills can be expected to be better than their speaking skills. If on the other hand, learners have received an input predominantly based on spoken language and oral interaction, their oral skills can be expected to be their strongest ones. In the present study, the learner profiles based

on test-results will be described and analysed, while the background factors that are hidden beyond the scores will be left for future research.

2.5 The Swedish Grading System of Modern Languages in Relation to the CEFR Levels

The linkage between the CEFR scales and the Swedish grading system makes it possible to relate and validate the oral and written grades given to the participating subjects in the Swedish system according to the CEFR levels. *Modern Languages*, which in the Swedish system is the term referring to any other foreign language but English, are eligible from the age of twelve or thirteen on in most Swedish schools. The foreign language is studied for a limited number of hours (i.e. two to three) per week. At the end of year Nine, the pupils are supposed to have accomplished the criteria of the *Modern Languages 2* course (*Moderna Språk 2*), which corresponds to the CEFR level A.2.1 (Bardel, Ericksson & Österberg 2019, for example, for an overview of foreign language education in Sweden). The relationship between the Swedish grades and the CEFR levels are described below.

The Swedish grading criteria are explicitly linked to the CEFR scales. This linkage has been shown in several analyses of the textual relationship between the Swedish curriculum and the levels of the CEFR, and also in studies in which the content and requirement levels of national tests were examined (Ericksson & Pakula 2017 and Erickson 2017 for details). The Swedish grading system consists of a six-point scale (A-F), in which F represent a non-pass, E a minimal Pass and A the highest grade level. Sample analyses have confirmed textual analyses, namely that a minimum requirement to be approved (i.e. to receive grade E) in *Moderna Språk 2* is the CEFR level A2.1. However, as pointed out by Ericksson & Pakula (2017), the question of how this relationship manifests itself in students' test results in the national tests of Modern Languages has not been investigated. For English, Borger (2018) found an overall correlation between the CEFR level stated and the Swedish grade E for the tested level B2, but no similar comparison has been made for Spanish. ESCL (2011) reported that only 14% of the Swedish participating subjects reached level A2.1, the target level to be attained after year Nine according to the Swedish syllabus. In order to compare the ratings of subjects' performances across different systems and thereby add further reliability to our results, one third of the total amount of data was evaluated according to CEFR standards.

3 The Study

3.1 Objectives

This study was carried out in a Swedish instructional setting, and the subjects tested were all L2-learners of Spanish with Swedish as the L1. The Research Questions (RQs) were as follows

:

1) What productive skill, speaking or writing, is the strongest one in the learner-performance of the subjects tested?

2) Is there any difference in the distribution of the Swedish grades (A-F) between these two skills in the group tested?

3) How do the grades awarded to the subjects in the Swedish grading system relate to the CEFR-levels?

On the basis of these premises, factors related to L2-learning in instructional settings will be discussed as possible explanations to the results, and suggestions for further research will then be outlined.

3.2 Methodological Concerns

For the present study, a quantitative approach was employed. An oral and a written task were used. The oral task was adapted from a national test design for a paired (or group) speaking test with peer-to-peer interaction, whereas the written task was based on a national written-production test. The level tested was that of *Moderna Språk 2,* as described in Section 2.5. The test tasks thus build on standardized models, related to the Swedish grading system, and thereby indirectly to the CEFR scales. This set-up also enables comparisons with results from other studies in which tested the same level was tested, such as ESLC (2011).

We are not oblivious of the fact that the national tests focus on the product rather than the process, and that these more process-oriented objectives of the curriculum therefore must be considered outside the scope of the test. Still, the national tests are used to support teachers' holistic assessment and final grading. The tests have a supportive role for the evaluation of pupils' knowledge in relation to the curricular goals and serve as a tool for securing national equivalence in the judgements of pupils' levels of knowledge in relation to the grading system (Ericksson & Börjesson 2001. Utbildningsdepartementet 2017: 22, Borger 2018: 18). In this sense, the summative test results measured in the national test format are indicators of grades, although the grades of the pupils are not solely based on the Swedish national test results, but also on formative assessments (assessments *for* learning and not *of* learning).

4 Research Design and Procedure

4.1 Informed Consent

The study design follows the Law on Ethical Review and Good Research Practice as indicated by the Swedish Centre for Research Ethics (http://www.codex.vr.se/en/manniska2.shtml; 20-07-2020). The participants' informed agreement was obtained prior to testing, with all participants signing a consent form. The subjects were in-

formed about the overall research plan, the aim of the research, the methods to be used, the fact that participation was voluntary and anonymous, and that they had the right to cease participation at any time.

4.2 Subjects

One of the general risks in data collection that may create bias in the results is that only the most interested and motivated learners volunteer to participate. In order to avoid the risk that the likeliness to participate would be linked to the outcome, and in order to control for background variables such as more or less study motivated learners, pupils from three different programs were recruited (Natural Sciences, Economics, and Social Sciences). Data collection was carried out as part of the ordinary teaching and did not require any extra time or extra effort of the participants. Out of a total of 109 students, 97 students participated in both the written and the spoken task (those subjects who only participated in one of the two tasks were removed from the study). Out of these 97 students, 90 agreed to participate in the study. When the tests the present study is based on took place, the 90 participants had just started their studies at Upper Secondary School. The subjects came from 20 different Secondary Schools located in rural as well as in urban areas of the region of Västerbotten in Northern Sweden. Out of the participating 90 subjects, 23 studied in the program of Social sciences, 31 in the Economy Program and 36 in the program of Natural Sciences. 70% of the subjects were girls and 30% were boys – a proportion which more or less reflects the number of male and female students of Spanish in Swedish schools. The data collected include a range of different study backgrounds. The distribution of boys and girls differs from the national total (53% girls and 46% boys, according to data from the Swedish Board of Education (Skolverket 2019)), with a higher number of girls being present in our data. Statistics from the Swedish Board of Education (Skolverket 2019) indicate a higher mean grade for girls than for boys in Modern Languages at a national level after finishing Secondary School, which implies that the mean grade of the present study probably is slightly higher than it would have been if the data had contained a more equal distribution of boys and girls.

4.3 Design of the Written and Oral Assignments

The oral interaction test consisted of a paired conversation during which pupils distributed in pairs discussed a given topic related to the description of a picture. The guidelines for test-takers of this task were similar to those used in the Swedish national test, and the topic was similar, although slightly modified. The written assignment consisted of the writing of an email directed to a Spanish fictitious friend. The oral and written tests will not be presented in detail, in order to protect test secrecy. For the oral task, audio recordings were used, an arrangement which eliminated the potential influence from subjects' ability to use alternative strategies, such as gaze or body language. The subjects were paired girl+boy[2] whenever possible, but since 70% of the data were produced by girls, this was not possible all through the way. Apart from this

[2] This arrangement is common in oral testing in general and in Swedish national oral tests in particular in order to make it easier to distinguish the different voices and thereby to facilitate the evaluators' work.

general guideline, the subjects were allowed to create pairs themselves, since we wanted to create an environment which be as calm and safe as possible for the test-takers. In the few cases in which voice recognition was a problem, the author (who has expertise in the phonetic domain) supported the raters in distinguishing the test-takers' voices.

4.4 Rating Process

In the EACEA report (2015), two thirds of the countries (22 out of 36) included had the written parts of the national tests (test of receptive skills and writing) externally marked. The report points out that in countries where speaking is assessed, this is often done internally, which makes the rating process for speaking different from the one for the other skills tested. In Sweden, external rating is recommended for all skills, and it is furthermore suggested that two or more external teachers collaborate in the evaluation process of pupils' assessments (Utbildningsdepartementet 2017, Skolverket 2020). The rating of the present study was carried out by external teachers who were not the ordinary teacher of any of the groups. Altogether, four raters were involved, one first and one second rater of each system, i.e. the Swedish grading system and the CEFR levels. It is worth pointing out that all four raters involved had more than ten years of experience of inter-colleague co-rating, the Swedish raters according to Swedish national guidelines and the Spanish raters according to the CEFR descriptors. Based on the Classical Test Theory[3], we therefore assume that any systematic rating error related to teacher experience or performance criteria did not occur.

The four raters involved in the evaluation of test results were certified and experienced teachers of Spanish as an L2, with teaching / rating experience ranging from 20 to 35 years, with a mean of 29 years. The data were evaluated according to the criteria of

a) the guidelines for the grading of the Swedish national test for *Moderna Språk 2,* presented by the Swedish Board of Education, and

b) the guidelines used for issuing the internationally recognised official exams, *Diploma de Español Lengua Extranjera*[4]. These qualifications strictly follow the levels and sublevels described in the CEFR: A1.1-A1.2, A2.1-A2.2 etc.

The grading of the assessments according to the Swedish grading system was carried out prior to the evaluation according to the CEFR descriptors. The guiding criterion for the selection of the data to be evaluated according to the CEFR descriptors was that the whole grading scale – from grade A (the top note) to grade F (fail) –, as defined by

[3] According to the Classical Test Theory (CTT), systematic errors may occur in cases in which raters are inadequately trained or where the respective evaluative performance criteria are inadequately specified (e.g. Henning 1996: 54 and Gulliksen 1987).

[4] The Diplomas in Spanish as a Foreign Language (*Diplomas de Español como Lengua Extranjera*; DELE) are officially and internationally accredited qualifications issued by the *Instituto Cervantes* on behalf of the Ministry of Education and Science of Spain according to the CEFR descriptors.

the Swedish raters, should be represented. According to this criterion, the productions were randomly selected from each one of the three participating study programs (10 from each one). The data evaluated according to the CEFR descriptors are thus not representative of the distribution of grades in the whole sample.

Altogether, 180 productions (90 written and 90 spoken) were rated by the Swedish first rater according to the Swedish grading system. Out of these, one third were evaluated by the Spanish first rater according to the CEFR scales. A second rating was then carried out by the respective second raters in order to check for inter-rater agreement (Section 5.1). The second evaluation process was carried out altogether independently from the first one. This procedure followed the guidelines proposed by the Swedish Ministry of Education in order to enhance equity in assessments, which suggests that a minimum of two independent raters are involved in the rating process (Svenska utbildningsdepartementet 2017). Below, the results from the comparisons are presented in descriptive and correlational statistics.

5 Results and Analysis

5.1 Intra- and Inter-Rater Reliability

Intra- and inter-judge reliability was calculated for both the Swedish and the Spanish raters. Spearman's rank order correlation was computed in order to assess statistical relationships based on the rank order of the ordinal data. The significance level was set to $p < 0.05$. According to McNamara (2000: 580) 0.90 is equivalent to a high level of agreement, while the acceptable level is set to 0.7. Intra-rater reliability was checked for the two first raters on one third of the complete data set (30 written and 30 spoken productions) with the aim to check if the judgements made by the main raters were stable (see also Good et al. 2015). In order to avoid bias from the first rating, the re-evaluation of the data by the same judge was carried out one month after the first rating. The results for both raters show a high intra-rater agreement: for the Swedish rater, the results were calculated to 0.979 (Spearman's rho) for the Oral data ($p < 0.001$), and 0.971 (Spearman's rho) for the written data ($p < 0.001$). The Spanish rater's first and second judgements of the writing task (30 assessments) was computed to 0.926, $p < 0.001$ (Spearman's rho). For the speaking test (30 productions), the intra-rater agreement of the Spanish rater was found to be a remarkable 100%.

The second rating carried out by the first raters was used as a reference point for the inter-rater comparisons. 20% of the whole data set, i.e. 36 assessments (18 written and 18 oral productions), were evaluated by the second Swedish rater according to the Swedish grading system. These grades were compared with those awarded to the subjects by the first rater. Spearman's rho was calculated to 0.989 for the Oral ($p < 0.001$) and 0.927 ($p < 0.001$) for the written performances. Inter-rater reliability was also checked for the Spanish raters, where 30% of one third of the data, i.e. 18 assessments (9 written and 9 oral productions), were rated by a second Spanish rater. Inter-judge agreement between the two Spanish raters was surprisingly found to be 100% for both written and oral assignments. The correlations are in agreement with, for example, Ericksson (2009), where Spearman's rho was calculated to >0.9 for raters who had previously worked together with co-rating. Other studies, for example

Borger (2018), which included judges who had not worked together before the rating, showed lower degrees of inter-rater agreement, ranging from 0.59-0.95 (Borger 2014: 77).

Considering that there was no rater-specific training for the present task and that the raters had not been allowed to communicate with each other, the consistency in marking appears to be robust. In the few cases where the Swedish raters had disagreed on an assessed grade (the differences in judgements were never bigger than one grading step), a discussion was held until consensus was reached. The comparisons between the Swedish grading system and the CEFR scales included grades of each system finally agreed upon. Based on the high statistical correlation levels found for both rating pairs, the tendencies observed should be reliable across both the Swedish and the CEFR grading systems, a weakness of the study being that only two raters were involved in each system.

5.2 Results for Written and Oral Performance in the Swedish Grading System

In order to answer Research Question 1 (*What productive skill, speaking or writing, is the strongest in the learner-performance of the tested subjects?*), results were computed for the 90 oral and 90 written grades awarded in the Swedish system, and learner profiles were analysed. The results show that the dominant profile is the learner who has stronger writing than speaking skills. The average percentage of test-takers that were awarded a higher grade in the writing task was 48%, while only 13% received a higher grade in speaking than in writing. In 38% of the cases, the test-takers had a balanced profile with the same grade in both tasks:

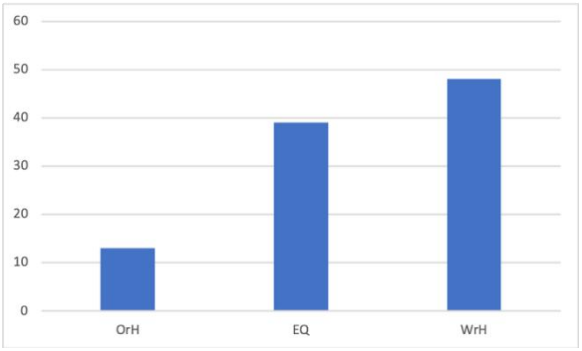

Figure 1: Learner Profiles in Productive Skills

(Distribution of cases where test-takers received the higher grade in oral proficiency (OrH), equal grades in the two skills (EQ), and a higher grade in writing proficiency (WrH))

As regards Research Question 2 (*Is there any difference in the distribution of the Swedish grades, A-F, between these two skills in the tested group?*), a matched samples test (Wilcoxon Signed Ranks Test) revealed significant differences between the oral and writing grades at the $p<0.001$ level. Figures 2 and 3 show the distribution of grades in oral and written performance according to the Swedish grading system. The most salient results are that the top grades (A and B) were obtained by the subjects to a higher extent in writing than in speaking, and that a higher number received

grade F (fail) in oral than in written performance (26.7 compared to 15.6%). The third highest grade (C) is also awarded to more subjects in writing than in speaking:

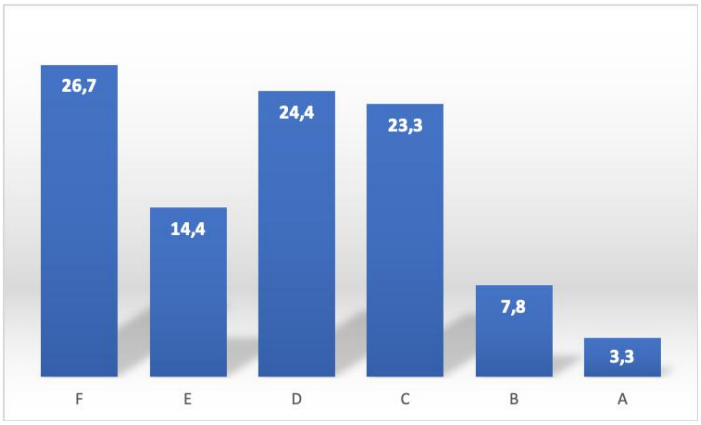

Figure 2: Oral Grades (90 participants) – Swedish Grading System

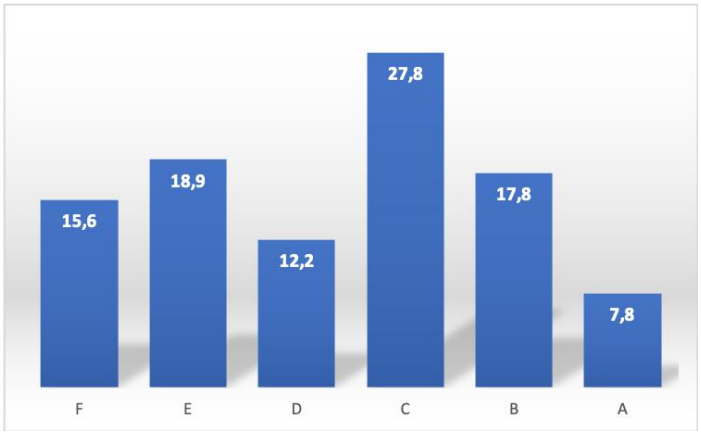

Figure 3: Writing Grades (90 Participants) – Swedish Grading System

5.3 Comparison of Grades Achieved in the Swedish Aystem and the CEFR Scales

The result of Research Question 3 (*How do the grades awarded to the subjects in the Swedish grading system relate to the CEFR-levels?*) is presented below in Figures 4 to 7. Here, the two rating systems – and especially the oral productions – seem to coincide to a great extent. This result may even be more surprising for the speaking task, since speech, due to its ephemeral nature, is more difficult to evaluate (Isaacs 2016: 8). As shown in the figures, the sample evaluated in both systems seems to be fairly normally distributed. In Figures 4 and 6, the CEFR levels obtained in the production tasks are presented according to the CEFR scales. Figures 5 and 7 show the grades awarded to the same productions in the Swedish grading system:

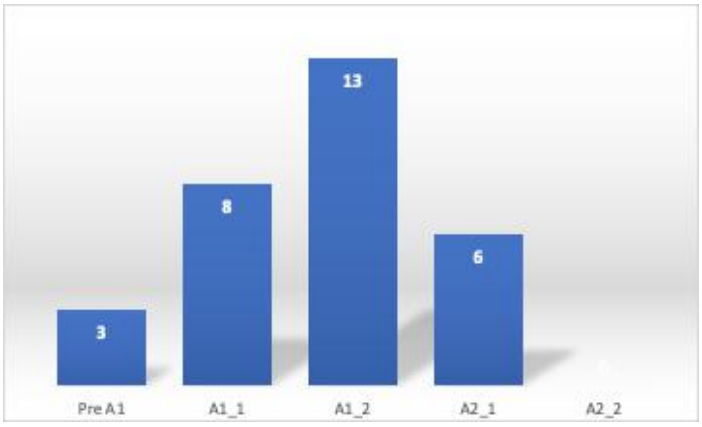

Figure 4. CEFR Ratings of Oral Performance: 30 Productions

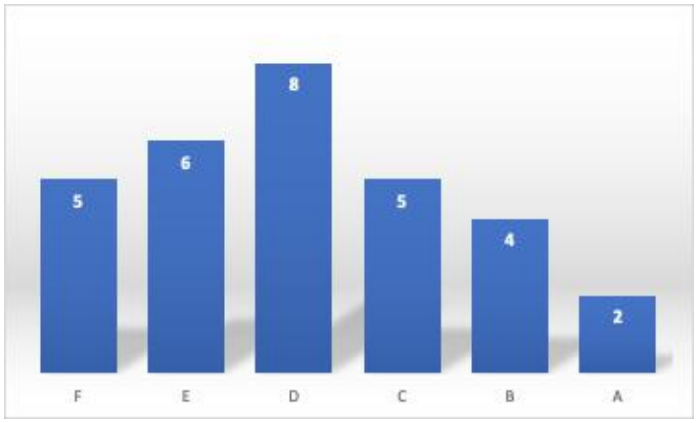

Figure 5. Swedish Ratings of Oral Performance: 30 Productions

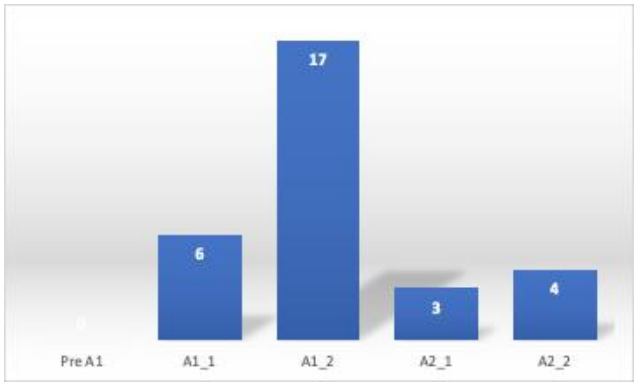

Figure 6. CEFR Ratings of Writing Performance: 30 Productions

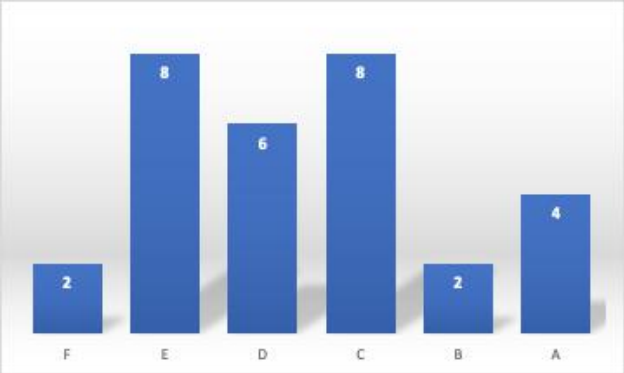

Figure 7. Swedish Ratings of Writing Performance: 30 Productions

In the field of oral performance, six productions were classified as level A2.1 (Figure 4), which represented the highest level among the subjects tested. If we compare these results with the Swedish gradings, Figure 5 shows that four subjects were awarded the second highest grade (B) and two subjects were awarded an A (six productions in toto). Eleven performances were classified as Pre-A1 or A1.1 according to the CEFR scales. This number corresponds to the number of fails, viz. grade F (5 productions), and grade E (6 productions) in the Swedish system.

In writing, seven subjects were awarded the level A2.1 (3) and A2.2 (4) (Figure 6). Approximately the same number, i.e. six subjects received the highest grades A (4) and B (2) in the Swedish system (Figure 7). No one was rated PreA1 in writing, and only two subjects were awarded grade F. In both systems, then, the lowest level (F / Pre A1) was awarded to more speakers in the oral task than in the speaking task. The mid-grades E-C and level A1.2 accumulated the highest number of subjects both in writing and speaking. In both the writing and the oral task, the subjects who reached the levels A1.1-A1.2 and E-C were approximately the same number (23 CEFR and 22 Swedish grades in writing, and 21 CEFR and 19 Swedish grades in oral performance). The resulting agreement between the ratings in the two systems shown in Figures 4 to 7, however, does not clarify the interrelationship between the ratings at an individual level. In the following paragraphs, Figures 8 and 9 show how the grades awarded in the two systems related to each subject across the systems.

The comparison of each individual's rating in the two systems is presented in three-dimensional graphs, in which the two axes represent one of the two grading scales each (Figures 8 and 9). A cross-comparison of the ratings shows that the Swedish grades are awarded at a slightly higher level in the CEFR scales in writing than in oral performance. Despite smaller differences in the ratings, as described below, the overall agreement between the systems was substantial. In Table 1, the relationships found in the comparison of Figures 8 to 9 are summarised:

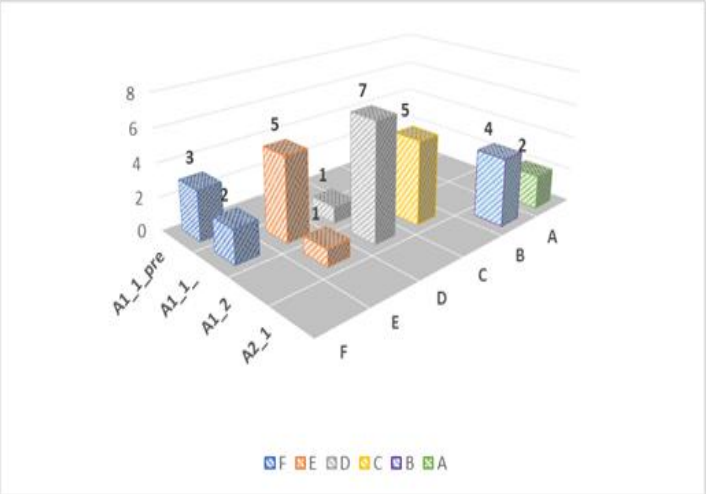

Figure 8: Distribution of grades in 30 oral assessments according to the CEFR scales and the Swedish grade descriptors

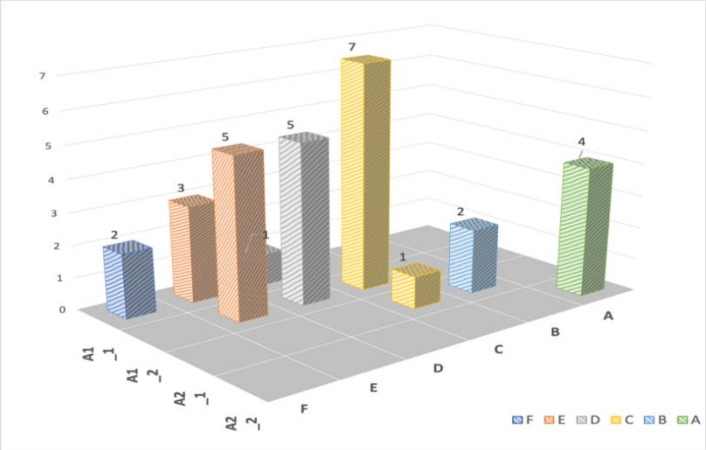

Figure 9: Distribution of grades in 30 written assessments according to the CEFR scales and the Swedish grade descriptors

Oral performance	PreA1/A1.1= F, E	A1.2= E (1), D, C	A2.1= B, A	
Written performance	A1.1= F, E, D (1)	A1.2 = E, D, C	A2.1= C (1), B	A2.2= A.

Table 1. Correlations Between the CEFR Scales and the Swedish Grades.

In the field of oral performance, three of the five speakers with grade F were awarded Pre-A1.1 by the Spanish raters, and in writing, two speakers with grade F were classified as A1.1. All but one oral E-grade was classified as A1.1 in the CEFR scales (the outlying E being assigned level A1.2, together with assessments that were classified as D or C). In writing, the E-grades were classified as CEFR level A1.1 (3) or A1.2 (5). The subjects who reached level A1.2 in the CEFR scales had been awarded the grades E, D or C in the Swedish system both in writing and in oral

performance. Finally, the subjects with the Swedish top grades (A or B) in both speaking and writing were all but one (a C grade in writing) awarded level A2.1-A2.2 in the CEFR scales. A closer inspection of the single C-grade in writing showed that this written production had initially been awarded a B by the Swedish first rater and a high C by the second rater. The raters had finally agreed to a high C.

6 Conclusions and Directions for Future Research

The results show that the investigated group, 15-16 year-old Swedish L2-learners of Spanish in an instructional setting, performed generally better in writing than in speaking. While 48% of the learners were awarded a higher grade in writing than in speaking, only 13% had a higher grade in oral performance. The highest grades (A and B) were more frequent in writing and a higher number of fails (grade F), were obtained by the subjects in speaking than in writing. The correlations between the grades / levels obtained by the subjects in the CEFR scales and the Swedish grading system add further reliability to the results. The overall results show that subjects achieved a higher level in writing than in speaking in both systems, which is interesting since oral performance might run the risk of being marked more generously (Smith 2016).

An additional finding is that only the top grades (A and B) seemed to reach the target CEFR level A2.1, also described as the threshold level for the lowest pass grade (E) in the Swedish syllabus. In the present comparison, none of the productions which were classified as grade E, neither in writing nor in speaking, qualified for level A2.1 according to the CEFR scales. If translated to the finally agreed grades in the Swedish grading system, these results, based on the assumption that only the grades A or B reached A2.1, indicate that only around 11.1% of the oral productions and 25.6% of the speaking productions were likely to achieve this level. This amount comes fairly close to the percentage of participants who reached level A2.1 or higher in the ESLC (2011), viz.14.0%.

As pointed out by e.g. Smith (2016), Weissberg (2000) and Muñoz (2014), many factors are probably involved when the “spikiness” of learner profiles is to be explained, such as learner motivation, cognitive maturity, mode preferences or the degree of exposure to the target language. The types and amount of in- and output that L2 learners have access to in instructional settings, have turned out to be more important predictors of learners’ progression than, for example, their starting age (Muñoz 2014). A possible dominance of writing-based activities in the classroom and a low degree of extramural exposure are suggested as explanatory factors to the results found in this study. However, further studies are needed to support this claim. In a follow-up study, I aim at investigating the type of input and the type of instruction that the subjects included in the present study have been exposed to, inside and outside the classroom. Classroom activities as well as activities that involve exposure to the target language outside the classroom will be related to the current findings in order to investigate underlying explanations for why the majority of the learners are more skilled in writing than in speaking.

Acknowledgements

First of all, I would like to thank the teachers involved in the ratings: they all worked without remuneration, just because they were interested in the subject: the first raters Per Löfstrand, Auxiliadora García García, and the second raters Mirja Wallström and Maria Teresa Corbacho Moncada. Without them, this study would not have been possible to carry out. I also want to express my gratitude to Anders Steinvall for his indispensable help with the three-dimensional graphs in Figures 8 and 9, Mathias Lundström for his helpful suggestions regarding the statistics. Finally, I would like to thank Lars Fant for his valuable comments on the final draft of the manuscript.

Funding

The project received financial support from Umeå School of Education (USE), Umeå University, Sweden.

References

Aronsson, Berit (2014). Prosody in the Foreign Language Classroom, Always Present Rarely Practised? *Journal of Linguistics and Language Teaching* 5 / 2: 207-224.

Aronsson, Berit (2016). Hur undervisar vi i de främmande språkens prosodi? Lingua 01: 42-47.

Baba, Kyoko, Yuri Takimoto & Miho Yokoshi (2013). Relationship between second language speaking and writing skills and modality preference of university EFL students. Bulletin of Department of English, Kinjo Gakuin University,金城学院大学論集 社会科学編 第 10 巻第 10 / 1, 56-68.
(https://researchmap.jp/k_baba/published_papers; 30-05-2020).

Bardel, Camilla, Gudrun Ericksson, Rakel Österberg (2019). Learning, teaching and assessment of second foreign languages in Swedish lower secondary school – dilemmas and prospects. *Apples – Journal of Applied Language Studies* 13 / 1, 7–26.

Bonnet, Gérard (ed.) (1998). *The effectiveness of the teaching of English in the European Union: report and background documents of the colloquium held in Paris on October 20th and 21st 1997*. Paris: Ministère de l'éducation nationale.

Bonnet, Gérard (ed.) (2004). *The assessment of pupils' skills in English in eight European countries 2002: A European Project.* Paris: Ministère de l'éducation nationale.

Borger, Linda (2014). *Looking Beyond Scores: A Study of Rater Orientations and Ratings of Speaking.* Licentiate thesis. Department of Education and Special Education, University of Gothenburg.

Borger, Linda (2018). *Investigating and validating Spoken Interactional Competence.* PhD. Diss. Department of Education and Special Education, University of Gothenburg.

Clark, Eve, V. (2003). 'Critical periods, time, and practice,' University of Pennsylvania *Working Papers in Linguistics* 9, 39–48.

Dimitrakopoulou, Georgia (2017). The Status of English and Other Foreign Languages in the Greek Educational Context. *The Warwick ELT online*, January 31.

Erickson, Gudrun & Lena Börjesson (2001). "Bedömning av språkfärdighet i nationella prov och bedömningsmaterial". In *Språkboken*, Rolf Ferm & Per Malmberg (eds.), 254-269. Stockholm: Myndigheten för skolutveckling.

Erickson, Gudrun (2017). *Holistic peer analyses of national tests in relation to the CEFR.* Presentation at EALTA CEFR Special Interest Group Meeting, London. (http://www.ealta. eu.org/resources.htm#Events; 30-05-2020).

Erickson, Gudrun & Heini-Marja Pakula (2017). Den gemensamma europeiska referensramen för språk: Lärande, undervisning, bedömning – ett nordiskt perspektiv [The Common European Framework of Reference for Languages – a Nordic Perspective]. *Acta Didactica Norge - nasjonalt tidsskrift for fagdidaktisk forsknings- og utviklingsarbeid* 11 / 3, 1–23. (https://www.journals.uio.no/index. php/adno; 30-05-2020).

Erickson, Gudrun (2009). Nationella prov i engelska - en studie av bedömarsamstämmighet [National tests of English - a study of rater agreement]. Stockholm, Sweden: The Swedish National Agency for Education.

European Commission (2012). *First European Survey on Language Competences: Final Report.* (https://www.researchgate.net/publication/262877352_First_European_Survey_on _Language_ Competences_Final_Report; 30-05-2020).

European Commission (2015). *Languages in Secondary Education – An Overview of National Tests in Europe 2014/15*. European Commission/EACEA/Eurydice 2015).

Ginther, April & Xun Yan (2018). Interpreting the relationships between TOEFL iBT scores and GPA: Language proficiency, policy, and profiles. In *Language Testing* 35 / 2, 271–295.

Good, Joy E., Dee M. Lance & Jacquie Rainey (2015). The effects of Morphological Awareness Training on Reading, Spelling, and Vocabulary Skills. Conversation Disorders Quarterly, 36 / 3, 142-151.

Gulliksen, Harold (1987). *The Theory of Mental Tests.* Hillsdale. NJ. Lawrence Erlbaum Associates.

Harris, Roy (2001). *Rethinking writing*. London: Continuum.

Henning, Grant (1996). Accounting for nonsystematic error in performance ratings. Language Testing, 13(1), 53-61. doi: 10.1177/026553229601300104.

Henriksen, Nicholas C., Kimberly Geeslin & Erik W. Willis (2010). The Development of L2 Spanish Intonation During a Study Abroad Immersion Program in León,

Spain: Global Contours and Final Boundary Movements. *Studies in Hispanic and Lusophone Linguistics* 3 / 1, 2-49.

Isaacs, Talia (2016). Assessing speaking. In Dina Tsagari & Jayanti Banerjee (eds.), *Handbook of second language assessment.* Berlin: DeGruyter Mouton, 131–146.

Kaulfers, Walter (1944). Wartime development in modern-language achievement testing. *Modern Language Journal* 28(2), 136-150.

Kroll, Barry (1981). Developmental relationships between speaking and writing. In Barry Kroll, & Roberta Vann (eds.), *Exploring speaking-writing relationships: Connections and contrasts*, 32-54. Urbana, IL: National, Council of Teachers of English.

Linell, Per (2005). *The Written Language Bias in Linguistics*. New York: Routledge.

McNamara, Tim (2000). *Language Testing.* Oxford, Oxford University press.

Muñoz Carmen (ed.) (2012). *Intensive Exposure Experiences in Second Language Learning.* Bristol: Multilingual Matters.

Muñoz Carmen (2014). Contrasting Effects of Starting Age and Input on the Oral Performance of For- eign Language Learners. *Applied Linguistics* 35 / 4, 463–482.

Oikonomidis, Agapios (2003). The impact of English in Greece. *English Today* 19 / 2, 55-61.

Olsson, E. (2011). ‘‘Everything I read on the Internet is in English’’ – On the impact of extramural English on Swedish 16-year-old pupils’ writing proficiency. Lic, Gothenburg: University of Gothen- burg.

Pérez Vidal, Carmen, María Juan Garau, Joan C. Mora, & Margarida Valls Ferrer (2012). Oral and Written Development in Formal Instruction and Study Abroad: Differential Effects of Learning Con- text. In Muñoz C. (ed.). *Intensive Exposure Experiences in Second Language Learning.* Bristol: Multilingual Matters.

Sandlund, Erica & Pia Sundqvist (2011). Managing task-related trouble in L2 Oral Proficiency tests: Contrasting interaction data and rater assessment. *Novitas Royal (Research on Youth and Language),* 5 / 1, 91-120.

Smith, Karyn (2016). How spiky can a spiky profile be? In T. Pattison (ed.) *IATEFL 2015 Manchester Conference Selections*. Faversham, Kent: IATEFL, 190-192.

Spolsky, Bernard (1995). *Measured words: The development of objective language testing*. Oxford: Oxford University Press.

Sundqvist, Pia (2009). *Extramural English matters: Out-of-school English and its impact on Swedish ninth graders’ oral proficiency and vocabulary.* PhD, Karlstad University Studies: Karlstad.

Sundqvist, Pia, and Sylvén, L. Kerstin (2012). World of VocCraft: Computer games and Swedish learners’ L2 vocabulary. In Reinders, H. (ed.), *Computer games in language learning and teaching*. Basingstoke: Palgrave Macmillan, 189–208.

Sylvén, L. Kerstin (2004). *Teaching in English or English teaching? On the effects of content and language integrated learning on Swedish learners’ incidental vocabulary acquisition.* PhD, University of Gothenburg, Gothenburg.

Sylvén, L. Kerstin & Pia Sundqvist (2012). Gaming as extramural English L2 learning and L2 proficiency among young learners. *ReCALL, 24 / 3,* 302–321.

Skolverket [Swedish National Board of Education] (2004). *Engelska i åtta europeiska länder: En undersökning av ungdomars kunskaper och uppfattningar.* Stockholm: Skolverket. Skolverket (2012a). Swedish National Board of Education, "Internationella Språkstudien 2011", *Rapport 375.* Stockholm: Skolverket.

Skolverket [Swedish National Board of Education] (2012). *Bedömning av språklig kompetens – En studie av samstämmigheten mellan Internationella språkstudien 2011 och svenska styrdokument. Skolverkets aktuella analyser 2012* [Assessment of language competence – A study of the alignment between the international language study 2011 and Swedish curricula. Current analyses by the Swedish National Agency for Education 2012]. Stockholm: Skolverket.

Skolverket [Swedish National Board of Education] (2013). *Att tala eller inte tala spanska: En fördjupning av resultaten i spanska från Internationella språkstudien 2011* [To speak or not to speak Spanish: An in-depth study of the results of the European Language Study 2011]. Stockholm: Skolverket.

Skolverket [Swedish National Board of Education] (2019). *Slutbetyg i grundskolan*. Report Dnr: 5.1.1 -2019:1342.

Skolverket [Swedish National Board of Education] (2020). *Samsyn som redskap för ökad likvärdighet.* (https://www.skolverket.se/skolutveckling/forskning-och-utvarderingar/forskning/sambedomning- som-redskap-for-okad-likvardighet; 30-05-2020).

Sylvén, L. K., & Pia Sundqvist (2012). Gaming as extramural English L2 learning and L2 proficiency among young learners. *ReCALL,* 24 / 3, 302–321.

Utbildningsdepartementet [Swedish Ministry of Education and Research] (2017). *Nationella prov – rättvisa, likvärdiga, digitala*. [National tests - fair, equivalent and digital]. Government Proposition, publication number Prop. 2017/18:14]. Stockholm: Swedish Ministry of Education and Research.

Weissberg, Robert (2000). 'Developmental relationships in the acquisition of English syntax: writing vs.speech'. *Learning and Instruction*, 10, 37-53.

Weissberg, Robert (2006). 'Talking about writing: Cross-Modality research and Second Language Speaking/Writing Connections. In Paul Kei Matsuda, Tony Silva (eds.), *Second Language Writing Research, Perspectives on the Process of Knowledge Construction*, Routledge: New York and London, 93-104.

Author:

Berit Aronsson, PhD

Associate professor of Spanish
Umeå University
Department of Language Studies
Humanisthuset
Umeå universitet
901 87 Umeå
Sweden
E-mail: berit.aronsson@umu.se