Organisation for Economic Co-operation and Development

EDU/WKP(2023)13

Unclassified **English text only**

28 July 2023

DIRECTORATE FOR EDUCATION AND SKILLS

Does English instruction teach more reading than listening skills? Evidence from 15 European education systems

OECD Education Working Paper No. 298

Gabriele Marconi, Agence pour le développement de l'emploi, Luxembourg.

This working paper has been authorised by Andreas Schleicher, Director of the Directorate for Education and Skills, OECD.

Contact:
Gabriele Marconi, (gabriele.marconi@adem.etat.lu)
Catalina Covacevich (catalina.covacevich@oecd.org)

JT03523758

OECD EDUCATION WORKING PAPERS SERIES

---

www.oecd.org/edu/workingpapers

Acknowledgements

This research greatly benefitted from discussions with Francesco Avvisati, Carla Campos Cascales, Catalina Covacevich, Tue Halgreen and Jimena Vargas. Sanneke Schouwstra and Ivailo Partchev helped with explanations on the data and needed methodology / packages for the analysis.

Abstract

This study investigates whether English formal instruction and a number of teaching practices are more strongly associated with reading or listening English skills, using data from a large-scale assessment of English skills among 14- and 15-year-olds in 15 European education systems in 2012. The results indicate that the skill difference between reading and listening skills is positively associated with: more years spent learning English in school; more hours of current English instruction; and even indicators of quality of English instruction. In addition, the use of different teaching materials and the emphasis put on oral skills in the classroom are also associated with the difference between reading and listening skills. These results are based on a methodology developed specifically for this study, and they confirm the usefulness of separately measuring foreign language skills for policy analysis.

Table of contents

FIGURES

TABLES

Introduction

Governments use various policy levers to influence foreign language learning, and it is reasonable to assume that each of these levers is associated in a different way with different language skills (for example reading, listening or speaking). This assumption underlies major policy and statistical work. For example, PISA countries and economies decided that the PISA 2025 Foreign Language Assessment should produce distinct measures for reading, listening and speaking, so that the statistical and policy analysis will be done separately for each skill (OECD, 2021[1]). This paper investigates this differential policy-skill association for English with a focus on reading and listening. English is the language tested in the first round of PISA's Foreign Language Assessment, and it is the most-demanded foreign language in the labour market, at least in Europe (Marconi, Vergolini and Borgonovi, 2023[2]). Reading and listening have the advantage of being two receptive skills with relatively comparable scoring methods (Jones et al., 2012[3]), and in addition they are both covered in the other, recent large-scale assessment on foreign languages (i.e. the European Survey of Language Competencies or ESLC, (European Commission, 2012[4])).

The literature on learning differences between foreign language reading and listening skills reflects a widespread belief that, across a great variety of school settings and education systems, formal foreign language (including English) instruction teaches better reading than listening skills. The famous statement by Oxford (1993[5]) that listening is the "neglected stepchild" of foreign language teaching has been supported by more recent studies in various ways: based on anecdotal experience (e.g. (Tschirner, 2016[6]; Bozorgian, 2012[7])); assumed from an observed lack of English listening skills (e.g. (Green and Bracciodieta, 2019[8])); distilled from reviewed literature (e.g. (Buttler, 2007[9]; Isaacs, 2012[10]; Gilakjani and Sabouri, 2016[11]); found from student questionnaires (e.g. (Berkleyen, 2007[12])); implied as a washback effect from the structure of English testing (e.g. (Amengual-Pizarro, 2009[13]; Cheng, 1997[14])); or simply stated as a known fact (e.g. (Miller, 2001[15])). Very few exceptions state the contrary, i.e. that classroom teaching nurtures listening more than reading. (Brumen, Cagran and Rixon, 2009[16]) underline the emphasis on foreign language listening skills in the specific case of Croatia; and (Spoden, Fleischer and Leucht, 2020[17]) argue that the students in the German state of Schleswig-Holstein learn more reading skills until the first stage of upper secondary education, but find that they learn more listening in the final stage of upper secondary education.

This paper contributes to the literature in a number of ways. First, the hypothesis that formal English instruction is more strongly associated with the learning of reading compared to listening skills has not been put to a formal, convincing test yet. This paper does this based on a large-scale, international dataset on foreign language skills. Second, this paper also tests a set of hypotheses concerning a closely related question: whether teaching practices (emphasis on written versus oral English, use of different teaching materials, use of English in the classroom) also affect the relative acquisition of reading compared to listening skills. By testing these hypotheses, this paper demonstrates that different policy levers can be associated with the learning of different language skills. Third, this paper presents a method to test the association of any explanatory variable with the difference between two foreign language skill measures derived from a large-scale educational assessment.

Testing the hypotheses of this study raises some methodological challenges, as it involves estimating the association of explanatory variables with a skill variable (reading), conditional on another skill variable (listening), in the setting of a large-scale survey (e.g. (Schofield et al., 2015[18]). This can lead to serious bias, because (a) the skill variables are

measured with a substantial measurement error, and (b) measurement error in a control variable can induce bias in all regression coefficients (Klepper, 1988[19]); (Greene, 2003[20]). In addition to the problem of measurement error, skill measures are usually scaled according to convention (e.g. based on a standardisation with an average of 500 and a standard deviation to 100 (OECD, 2019[21]), in a way that does not make them comparable with each other. The solution adopted in this paper is to use a measure of the difference between reading and listening skills as the dependent variable. This is one way of conditioning reading on listening skills, without incurring the problem of including a noisy skill measure among the independent variables. To make the difference between reading and listening meaningful, both reading and listening skills are anchored to the Common European Framework of Reference (CEFR) (Council of Europe, 2001[22]; Council of Europe, 2018[23]). Finally, to correctly account for the measurement error, plausible values of this difference are drawn from the estimated bivariate normal distribution of reading and listening skills (e.g. (OECD, 2021[24])).

The results show that formal English instruction is associated with a more positive difference between students' reading and listening skills (the coefficients are significant even though not very large). This result holds for various measures of English instruction: the number of years a student has been learning English at school, the hours of English lessons currently attended by the student, and even proxies for the quality of the lessons (the use of English in the classroom or the emotional engagement of students). The evidence also confirms that teaching materials and the emphasis put by the teacher on written and oral English are associated with a change in the difference between reading and listening skills.

By using large-scale assessment data, this study addresses the problem on non-probability or non-diverse samples, a major limitation of much literature in applied linguistics (Sudina, 2021[25]; Andringa and Godfroid, 2020[26]). However, since the data were collected in 2012, much has changed in education, in particular due to the Covid-19 pandemics (Marshall, Pressley and Love, 2022[27]; Schleicher, 2022[28]). In addition, all the education systems covered by the survey were in the same continent (Europe). It will be of great interest to investigate the same hypotheses with more countries and more recent data when the results of the PISA 2025 Foreign Language Assessment will be out.

The research hypotheses investigated in this paper are derived in the next section. The third section provides a discussion of the main methodological challenges encountered in the design of the empirical methods. The fourth section describes the data source, i.e. the ESLC (Jones et al., 2012[3]; European Commission, 2012[4]). The fifth section describes in detail the generation of the dependent variable, while the following one discusses the explanatory and control variables used in the paper. The final two sections summarise the empirical results and draw some conclusions.

Research hypotheses

English in the classroom in "default mode": why the emphasis on reading

From a historical point of view, the purposeful teaching of listening has developed after that of reading and writing (Goh, 2008[29]). The learning of vocabulary and grammar and the practice of written translation has been the backbone of foreign language teaching throughout several centuries and in different context such as China (Zhou, Hiver and Al-Hoorie, 2021[30]), Japan (Hino, 1988[31]; Hosoki and Yukiko, 2011[32]); and Europe (Richards and Rodgers, 2001[33]). In Europe, this approach became dominant after the Middle Ages, when the rise of vernacular languages (e.g. English, French) reduced the

importance of Latin (the most-often taught second language at the time) as a spoken language. As a result, second language teaching shifted towards the analysis of grammar and the reading of the classics of Latin literature (Richards and Rodgers, 2001[33]).

With traditional teaching approaches favouring grammar and written competencies, the possibility to teach English listening skills at all was still a matter of scholastic debate in the 1950s (Brown, 1954[34]). The general view for much of the past century have been that the learning of listening would follow almost automatically from the teaching of speaking and reading skills, while scholars have by now agreed that an active emphasis on developing listening skills is needed (Goh, 2008[29]).

In the academic field of foreign language teaching, more emphasis on listening was brought forward by communicative language teaching (CLT). CLT denotes various conceptual approaches to teaching, with a common focus of putting the teaching of all communicative skills at the centre of foreign language teaching (Whitley, 1993[35]; Goh, 2008[29]; Celce-Murcia, 2014[36]). However, theory does not always correspond to practice and many teaching contexts are probably still oriented towards traditional methods (Marconi et al., 2020[37]; Graves and Garton, 2017[38]; Goh, 2008[29]). This could mean that many education systems are still in a "default mode", perpetuating an inherited emphasis on teaching more reading than listening skill, because they are unable to implement fully the new teaching paradigms advanced by pedagogical theorists. While this paper investigates this hypothesis for 2012 (the year of the ESLC data collection), new data from the PISA 2025 Foreign Language Assessment will make it possible to see if this "default mode" was shaken off in post-Covid education systems.

Hypothesis: Classroom teaching favours reading to listening

If classroom teaching is, on average across education systems, still more focused on teaching reading than listening skills, then students spending more time in the English classroom should be relatively better at reading (compared to listening) than other students. In other words, more classroom teaching should correspond to a larger reading-listening difference. The data used in this paper contain measures of the number of years in which students studied English (an indicator called Duration of English learning by (Jones et al., 2012[3])) and on the number of English classroom hours in the current year (English lesson time a week in (Jones et al., 2012[3])). With regard to these variables, the following sub-hypotheses should hold:

- H1a. Duration of English learning at school is positively associated to the reading-listening skill difference.
- H1b. English lesson time a week is positively associated to the reading-listening skill difference.

A robustness test: Learning outside the classroom does not favour reading (compared to listening) in the same way as classroom learning

The testing of H1 could lose its validity if serious methodological or data issues went undetected. For example, if reading skills were measured accurately and listening skills were not, this could result in a paradoxical situation in which all learning is empirically associated with a larger reading-listening gap. To exclude such hypothetical flaws, it is useful to test a complementary hypothesis: that learning outside the classroom is not necessarily associated with a larger reading-listening skill gap.

Besides being a useful robustness test, this hypothesis has some significance of its own. Face-to-face exposure to and use of foreign languages outside school, in particular, is expected to have a positive impact especially on listening, speaking and communication skills (Marconi et al., 2020[37]). Based on the data used for this paper, it is possible to build two different measures for face-to-face activities involving English use outside school: one for travelling abroad (number of trips abroad); and one for talking with English speakers closer to home (face-to-face use of English in home country).

Therefore, it is useful to establish the following research hypothesis (divided in two sub-hypotheses):

- H2a. The number of trips abroad is not positively associated with the reading-listening skill difference
- H2b. Face-to-face use of English in home country is not positively associated with the reading-listening skill difference

Hypothesis: Skill teaching emphasis is associated with skill learning

The argumentation above implies that education, in the "default mode", to some extent neglects listening skills compared to reading ones. This hypothesis can be tested based on questions on teaching activities inside the classroom, divided by the skill they focus on (see next sections). Given the limited number of items on this topic in the questionnaire, it is only possible to generate robust measures (based on at least three items) on emphasis on written English (including reading and writing skills) and oral English (including listening and speaking skills). This allows formulating two sub-hypotheses:

- H3a. Emphasis on teaching written English is positively associated with the reading-listening skill difference
- H3b. Emphasis on teaching oral English is negatively associated with the reading-listening skill difference

According to the argumentation above, the emphasis on teaching the different skills mediates the relationship between classroom teaching and skill learning. In other words, following the "default mode" argumentation outlined above, we would expect that accounting for the emphasis on reading vs listening in the model would eliminate the association between the current amount of classroom teaching (see H1b) and the reading-listening skill difference. If that does not happen, then there may be other factors in the classroom that favour reading compared to listening.

English in the classroom after the "communicative revolution"

Whitley (1993[35]) characterised the emergence of CLT as an "incomplete revolution". This idea, implying that there was a revolution of the conceptualisation of language teaching, which was not fully applied inside the classroom, has been echoed in various studies recently (Savignon, 2017[39]; OECD, 2021[1]). This raises the question of whether the partial application of new precepts in the classroom still favoured reading compared to listening. In other words, teachers may have been able to apply some of the new ideas of communicative teaching, but in a partial way that still favours reading compared to listening. This could have happened for various reasons, of which three are mentioned here as examples.

First, teachers' proficiency can be a blocking factor when implementing communicative language teaching, and a stream of literature (especially focussing on Eastern Asia)

document that teachers' English proficiency is higher in reading, compared to the other skills (Li, 1998[40]; Butler, 2004[41]; Ngoc Khoi, 2014[42]). Low teacher proficiency makes it difficult for teachers to engage students and to effectively implement CLT (Choi and Lee, 2016[43]; Chacón, 2005[44]).

Second, it could be relatively easier to find and use authentic teaching material for reading activities than for listening. For example, Ahmed (2017[45]) compiles a list of authentic materials, from which the only ones that can be used without the support of IT or of additional human resources are written materials like books and magazines. Probably because of this reason, the use of written authentic materials has been practiced even before the advent of CLT (Gilmore, 2007[46]). In addition, authentic spoken materials present some additional difficulties like the need to reproduce the social interaction of casual conversation (Gilmore, 2007[46]).

Third, even though some scholars have criticised the use of textbooks, they are likely to be one of the most important ways through which CLT can spread into educational practices (Whitley, 1993[35]). However, as Whitley (1993[35]) acknowledged, this could entrench some of the biases of traditional teaching methods towards reading and grammar. This view is echoed by Funk (2012[47]) who posits that textbooks have originally being designed to teach predominantly reading and grammar, and have been changing progressively only after the spread of CLT ideas.

Hypothesis: Student engagement and use of English in the classroom are positively associated with skill learning

Ideally, testing the hypothesis that teaching favours reading even while applying some CLT precepts would require data on the application of CLT in the classroom. The data used in this paper do not measure this directly. However, it is possible to build from the data two indicators that are both general proxies for quality of English teaching and key ingredients of the CLT approach: student engagement and the use of English in the classroom.

The data allow creating a measure of at least one dimension of engagement, emotional engagement (see section on explanatory and control variables), measured through a set of items similar to the one discussed in Zhou et al. (2021[30]) for the same sub-construct. Student engagement is "a state of heightened attention and involvement, in which participation is reflected […] in cognitive […] social, behavioural and affective dimensions" (Philp and Duchesne, 2016, p. 51[48]). Student engagement plays a very important role in enhancing student learning, and it takes a particularly important role in the field of language learning because it is an essential ingredient of CLT delivery (Hiver, Al-Hoorie and Mercer, 2021[49]; Whitley, 1993[35]).

Another measure potentially related to the implementation of CLT is students' reported use of English during English lessons (Jones et al., 2012[3]). The use of the foreign language in the classroom lies at the core of CLT, which focuses on teaching a language through its use (Chambers, 1991[50]). For this reason, in a similar way as engagement, the use of English in the classroom can be used as an imperfect proxy for both quality of instruction and the implementation of CLT.

For both indicators, it can be hypothesised that they increase the reading-listening skill difference if CLT is applied in a way that still favours reading skills. An alternative reasoning that leads to the same hypothesis is that, if the two indicators are considered as generic proxies for quality (Goldspink and Foster, 2013[51]; Chambers, 1991[50]), it can be hypothesised that the quality of English instruction would have a similar effect as its quantity (see Hypotheses H1a and H1b).

- H4a. Student emotional engagement in the English classroom is positively associated with the reading-listening skill difference.
- H4b. Students' reported use of English during English lessons is positively associated with the reading-listening skill difference.

Hypothesis: Different teaching materials are associated with learning different skills

The argumentation that teaching may favour reading even while partially applying CLT relies, in part, on considerations on the teaching materials. The data used in this paper allow building some measures on the use of teaching materials: use of ICT in the classroom and use of the textbook in the classroom. This leads to an additional hypothesis that can be tested:

- H5a. The use of ICT in the classroom is positively associated with the reading-listening skill difference.
- H5b. The use of the textbook in the classroom is positively associated with the reading-listening skill difference.

Methodological challenges

General methodology

The hypothesised statistical associations (see previous section) have been assessed by means of multivariate regression analysis, with a measure of the difference in reading-listening skills as the dependent variable and a list of explanatory and control variables as the independent variables. Following the standards in large-scale assessment data analysis (OECD, 2009[52]; Braun and von Davier, 2017[53]), standard errors are clustered at the school level and all estimates are based on a Jackknife estimator with five repetitions (one per skill plausible value – see also dependent variable section). Survey weights are adjusted so that their sum is equal for each country. This implies that the coefficients reported in this paper represent the average across education systems, a standard indicator in international reporting (e.g. (European Commission, 2012[4]; OECD, 2019[21]). The regression estimates have been implemented through R using the Survey package (Lumley, 2004[54]; 2021[55]).

Problems in building a measure of reading skills conditional on listening skills

The main research question of this study requires assessing if, for students that learned more English at school, the level of reading skills is particularly high, conditional on listening skills. This implies that two separate skill measures (for reading and listening) must be analysed jointly, and together with additional variables. When skills can be assumed to be fully measured (i.e., through an item list long enough to measure every relevant aspect of the skill) this is not a problem. For example, in research comparing listening and reading skills, Bozorgian (2012[7]) used results from the IELTS, a test lasting almost three hours; while Sok et al. (2021[56]) used other tests that they considered of sufficient quality. However, such comprehensive tests cannot be administered to a large-scale probability sample.

Large-scale assessment surveys can only administer limited sets of questions to students, due to their time and cognitive constraints. This results in skill estimates that suffer from measurement error. The institutions collecting the data usually generate plausible values for student skills, accounting for measurement error, as random draws from an individual

skill distribution. These estimates (also known as institutional plausible values) are meant to be used as dependent variables either in univariate analysis or in multivariate analysis, with indices derived from the background questionnaires as dependent variables (Junker, Schofield and Taylor, 2012[57]; Schofield et al., 2014[58]; OECD, 2009[52]; Braun and von Davier, 2017[53]).

At first glance, a simple approach to investigate the research questions of this paper could have been to use institutional plausible values as both dependent (reading) and control (listening) variables, while adding more explanatory and control variables as independent variables. Plausible values are devised to estimate correctly the standard error of a skill estimate by artificially inflating its variance. This approach is the standard for a variable used as dependent variable in a multivariate regression, but it can introduce serious bias if the plausible values are used as a control variable. In the context of this study, the inclusion among regressors of variables measured through plausible values seems indeed to lead to serious bias in the regression coefficients (see section Results).

The potential introduction of bias when using plausible values as control variables is known from the literature. Junker et al. (2012[57]) and Schofield et al. (2014[58]) discuss the issue of using institutional plausible values among the explanatory / control variables in a regression, concluding that this can lead to substantial bias in the regression coefficients. They recommend avoiding the use of institutional plausible values as conditioning variables, and generating new ones based on a customised model instead. However, even this may not clear the estimates from bias. A more general problem is that introducing measurement error in an independent variable causes all the coefficients from all independent variables to be biased in any direction (e.g. (Greene, 2003[20])). Empirical assessments of this bias (e.g. (Klepper, Kamlet and Frank, 1993[59]), and (Marconi, 2018[60])) show that it can be substantial and, indeed, that it can push any coefficient in both directions. Since plausible values include a large, artificially-induced error component, using them as co-variates makes it difficult to reach any conclusion on the association between explanatory and dependent variables (even though in some applications it can happen that the bias is not large, see (Braun and von Davier, 2017[53])).

The approach taken to address this problem consists of two steps (which are explained in detail in the section on the dependent variable): (1) estimating new plausible values for English reading and listening skills, which are defined on a comparable scale anchored to the CEFR; and (2) using the difference between the two as the dependent variable for the model, to avoid having plausible values on both the left- and the right-hand side of the estimation equation. This approach also has the advantage of providing an intuitive interpretation of the results, as the regression coefficients can be interpreted as a difference in CEFR levels between reading and listening.

Data

This study uses data from the ESLC, a survey on foreign language skills promoted by the European Commission and administered in 2012. The data from this survey can be considered of very high quality, due to the careful design of the instruments and to the high response rates from schools (93% of those contacted participated to the survey) and students (90%) (Jones et al., 2012[3]).

The ESLC survey tested students' skills in both paper and computer-based formats in English, French, Spanish, German or Italian in 16 European education systems: Bulgaria, Croatia, England, Estonia, Flemish Community of Belgium, France, French Community of Belgium, German Community of Belgium, Greece, Malta, Netherlands, Poland, Portugal, Slovenia, Spain and Sweden. A sample of around 1 500 students per language and per

education system took the survey, which consisted of a language test and a background questionnaire. The students taking the test were at the last year of lower secondary education (ISCED2) or the second year of upper secondary education (ISCED3). Two languages were tested in each education system; English was tested in all education systems except for England (Jones et al., 2012[3]; European Commission, 2012[4]).

For each language, the three skills of listening, reading and writing were tested separately, with each student being tested in two of the three skills (chosen at random). The length of the language tests was around 30 minutes per skill, implying one hour of language tests for each student. The language tests, based on state-of-the-art methodologies and technologies, were designed by University of Cambridge ESOL Examinations (Cambridge ESOL-English for speakers of other languages), Centre international d'études pédagogiques (CIEP), Goethe Institut, Università per Stranieri di Perugia and Universidad de Salamanca.

The fact that each student was tested in two out of the three skills implies that one third (around 500 per education system) of the students tested in English in each education system took both the reading and listening test. This group constitutes the sample used for this study. It consists of 6 343 students from the 15 education systems testing English skills, out of a total of 47 797 students that took the ESLC survey across all education systems and languages.

The availability of such high-quality data based on probability sampling should not be taken for granted in the field of language learning. The use of the high-quality data provided by the ESLC avoids the problem of non-probability samples (Sudina, 2021[25]) and of samples that are biased towards college students in a handful of countries (Andringa and Godfroid, 2020[26]). A further improvement in this direction will be made possible by the PISA 2025 Foreign Language Assessment (OECD, 2021[1]), which will assess students in more countries, including non-European ones.

Dependent variable

The dependent variable used for the multivariate regressions in this study is the difference between reading and listening proficiencies, which were estimated based on a Rasch model (Pan, 2018[61]; Linacre, 1999[62]) and anchored to the CEFR. This difference was expressed as a set of plausible values, which in turn consisted of the difference between plausible values for reading and listening proficiencies.

Estimating plausible values of student skill levels

The Rasch-model parameters of the estimates of reading and listening proficiencies have been obtained through maximum likelihood (Pan, 2018[61]), with each individual response treated as a logistic function of the individual test score and binary variables representing the item being answered[1]. The Rasch model was a natural choice because it was the same that was used to estimate the skill level in the ESLC survey dataset (European Commission, 2012[4]; Jones et al., 2012[3]). Plausible values were imputed as draws from the bivariate distribution of reading and listening proficiencies, estimated from two latent regressions of the previously obtained estimates (see e.g. (OECD, 2012[63]; 2021[24])) on all the variables

[1] Each student was assessed through one of six question booklets, with a set of common items across booklets. The parameters for students' scores in the Rasch model estimation were booklet-specific. This implies different proficiency estimates for students who answered correctly the same number of items, but from different booklets. In contrast, the model's item difficulty parameters were constrained to be the same for all booklets containing them, so that these items could serve as anchors.

included in the model. This suits the recommendation by Schofield et al. (2014[58]) to use a customised conditioning model.

Anchoring the difference between reading and listening skills to the CEFR

To compare the skill measures in listening and reading obtained through the procedure described above, it is necessary to map them onto a common scale. This common scale was provided by the CEFR framework. This framework considers each communicative skill as functional to achieve an overall communicative language competence: "communicative language competence is activated in the performance of the various language activities, involving reception, production, interaction or mediation (in particular interpreting or translating). Each of these types of activity is possible in relation to texts in oral or written form, or both" (Council of Europe, 2001, p. 14[22]). According to the framework, the CEFR levels are a useful scale "to segment the learning process for the purposes of curriculum design, qualifying examinations, etc." (Council of Europe, 2001, p. 17[22]). In practice, this means that the framework sets an equivalence between, say, level A2 in reading and level A2 in listening, at least in terms of curriculum design, qualifying examinations and contribution to overall language competence. This is widely accepted in language education, at least in Europe. The large part of European countries define the expected learning outcomes at different stages of secondary education in terms of CEFR levels, and these expected outcome levels are generally equal across the four skills for a given grade (Eurydice and Eurostat, 2012[64]; Eurydice, 2017[65]).

The tasks of the ESLC were anchored to the CEFR, meaning that each task had been classified by language experts at a given CEFR level (Jones et al., 2012[3]).[2] The Rasch model postulates that the probability of correctly answering a given question is equal to the logistic transformation of the difference between a respondent's ability and the question's difficulty (OECD, 2012[63]; Boone, 2016[66]; Linacre, 1999[62]). Therefore, the probability that a given student is proficient at any level is calculated as the probability of each student to complete correctly a task with the average estimated difficulty for that same level.

$$(1) \qquad \Pr\left(L_{i,k,s}=1\right)=\frac{e^{\theta_{i,s}-\bar{d}_{k,s}}}{1+e^{\theta_{i,s}-\bar{d}_{k,s}}}$$

Where, for each skill $s=\{reading, listening\}$: $L_{i,k}=\{0,1\}$ is the fact that student i (with a skill score θ_i) has at least the level of skill $k=\{A1,A2,B1,B2,C1\}$, and $\bar{d}_k$ is the average estimated difficulty across all tasks at level k.

Because of the adherence to the CEFR, the probability of attaining a given skill level is measured on a comparable scale across the two skills of reading and listening. For example, the difference $\Pr(L_{i,B1,reading}=1) - \Pr(L_{i,B1,listening}=1)$ has a meaningful interpretation: it is the difference in the probability that student i reaches the same level (B1) in reading and in listening.[3] The same probability differences can be calculated for each of the five skill levels. Therefore, an intuitive approach to quantify the reading-listening skill difference is

[2] The ESLC classified questions at the levels A1, A2, B1 and B2. In addition, the probability that a student is at level C1 has been defined in this study as the probability that the student can answer the most difficult question in the whole assessment.

[3] In most EU countries, this coincided with the expected learning outcomes for both reading and listening at the end of lower secondary education (Eurydice and Eurostat, 2012[64]).

to average or sum up the differences for each level. This paper uses the sum of these differences as the main dependent variable:

(2) $$ExpDiff_i = \sum_k \Pr(L_{i,k,reading} = 1) - \Pr(L_{i,k,listening} = 1)$$

The measure *ExpDiff*$_i$ can also be interpreted as the difference in the "expected value" of the CEFR level in reading and listening, with 1 standing for A1, 2 for A2, 3 for B1, 4 for B2 and 5 for C1 (and any non-integer number representing an intermediate position between levels). This interpretation has the advantage of being intuitive and easy to communicate, an important advantage for a scale (Braun and von Davier, 2017[53]). For example, a difference of –1 for a student is interpreted as a student being "1 CEFR level below in reading as compared to listening".

It is also important, however, not to assume linearity as an intrinsic property of this scale. The Council of Europe warns that "one should be careful about interpreting sets of levels and scales of language proficiency as if they were a linear measurement scale like a ruler. No existing scale or set of levels can claim to be linear in this way" (2001, p. 17[22]). More generally, as stressed by Braun and Von Davier (2017[53]), when generating skill measures there is not one single scale that can be considered as intrinsically better than the other possible ones: any monotone transformation of a skill scale is equally valid.

Explanatory and control variables

Some of the indicators required to test the research hypotheses have been made available by the consortium that prepared the data (Jones et al., 2012[3]). These measures, on the topic of English learning at school and the use of English in the classroom, are presented in Table 1. To deal with the skewed distribution of English lesson time a week (with some students reporting to attend 10 or more hours of lessons a week – (Jones et al., 2012[3])), the natural logarithm is used in the analysis. Using the log implies that the regression coefficients for this variable must be interpreted as the expected increase in the dependent variable for a 1% increase in the weekly number of hours.

Table 1. Ready-to-use indices used for testing hypothesis H1 and H4b

Index name (Jones et al., 2012[3])	**Summary description**	**Hypothesis**
Duration of English learning	Number of years during which students studied English at school	H1a
English lesson time a week (log-transformed)	Log of the weekly number of hours a week that students spend learning English in the classroom. This measure has been generated by multiplying the number of classes by their duration.	H1b
Students' reported use of English during English lessons	A measure of how much students speak English in the classroom (on a 0-4 scale derived from the underlying Likert scale)	H4b

For the other constructs needed to test the hypotheses put forward in this paper, it was necessary to build custom-made indices. The fact that new indices had to be calculated was expected, because the analysis by Jones et al. (2012[3]) and the European Commission (2012[4]) did not focus on individual skill differences. Fortunately, the student background questionnaire was extensive (about 45 minutes to complete, including many questions specific to English learning), providing with a good set of items to generate new indices. Table 2 reports the indices built, the related hypothesis, and the items used to construct them.

Following Jones et al. (2012[3]), the reliability of the latent variables was assessed based on Cronbach's alpha, corrected through the Spearman-Brown prophecy formula for 10 items.

As recommended by Taber (2017[67]), besides reporting on the widely followed criterion that α is around 0.7 or larger (Table 3), the full list of items used for the instruments is reported in Table 2. Again following Jones et al. (2012[3]), an index based on a set of items is generated as the first principal component extracted through principal component analysis. The proportion of data variance explained (Jolliffe, 2002[68]) by the first three principal components is also reported in Table 3. The main component always has a much larger explanatory power than the other components indicating that, for each set of items, the index represents the main underlying component.

Table 2. Items included in the custom-made indices generated to test the research hypotheses

Face-to-face use of English in home country	Emphasis on teaching written English	Emphasis on teaching oral English	Use of ICT in the classroom	Use of the textbook in the classroom	Emotional engagement
Do you, yourself, come into contact with English outside school in the following ways? Through English speaking tourists who visit the place where you live	How often do you do the following during English lessons? Learning to write in English	How often do you do the following during English lessons? Learning to speak in English	How often are the following resources used in your English lessons? Internet	How often are the following resources used in your English lessons? Textbook	My teacher of English is a good teacher
[Same question] Through English-speaking people who live in your place of residence?	[Same question] Learning to read English texts	[Same question] Learning to understand spoken English	[Same question] Computer programmes	How useful are your English textbooks, or is your English textbook, for the following? For learning English grammar	I get along with my teacher of English
How often do you speak English with people living in your place of residence?	How important are the following in order to get a good final grade for the subject of English? Writing English well	[Same question] Learning to pronounce English correctly	[Same question] Language lab	[Same question] For learning English words	My teacher of English makes an effort to make the lessons interesting for us
How often do you speak English with tourists?	[Same question] Reading English well	How important are the following in order to get a good final grade for the subject of English? Speaking English well			My teacher of English is helpful
		[Same question] Understanding spoken English well			I like my teacher of English
		[Same question] Pronouncing English correctly			My English lessons are interesting
					My English lessons are enjoyable
					My English lessons are good

Note: The Cronbach alpha has been adjusted through the Spearman-Brown prophecy formula for equivalence with a 10-item test.

Table 3. Adjusted Cronbach alpha and variance explained by the first three extracted components for each custom-made index

	Face-to-face use of English in home country	Emphasis on teaching written English	Emphasis on teaching oral English	Use of ICT in the classroom	Use of the textbook in the classroom	Emotional engagement
Adjusted Cronbach alpha	0.78	0.85	0.92	0.94	0.82	0.95
First component: variance explained	51%	54%	54%	74%	59%	69%
Second component: variance explained	21%	24%	25%	18%	28%	11%
Third component: variance explained	17%	12%	7%	8%	13%	5%

Note: The Cronbach alpha has been adjusted through the Spearman-Brown prophecy formula for equivalence with a 10-item test.

The only custom-built index that has not been included in Table 2 and Table 3 is Number of trips abroad, because it could be generated as a compound index through a meaningful arithmetical transformation of a set of items (Jones et al., 2012[3]). More precisely, the items used for the construct (Table 4) collect information about the number of trips abroad done by the student, so they could be summed up to generate a meaningful index representing the total number of trips abroad. Given the role of English as a lingua franca (Marconi et al., 2020[37]), trips to both English and non-English speaking countries have been included in the index.

Table 4. Items used for the index "number of trips abroad"

Item
How often did you go on a school trip to an English speaking country?
How often did you go on a school trip to another non-English speaking country?
How often did you go with your family to an English speaking country?
How often did you go with your family to a non-English speaking country?

The custom-built indices were calculated only for students who responded to at least one of the items used to construct the indices. For students who responded to at least one (but not to all) items, missing data points were filled through ratio imputation (de Waal, Pannekoek and Scholtus, 2011[69]), using as auxiliary variable the average across responded items.

In addition to the aforementioned variables, a set of indices related to students' background and use of English outside school was used as control variables in the regressions. These index were made available by the consortium preparing the data (Jones et al., 2012[3]), and they are presented in Table 5.

Table 5. Indices used as control variables

Index name (Jones et al., 2012[3])	**Summary description**
Gender	Female (0) or male (1)
Age	Years of age
Immigration background	Student has at least one parent born in the country of study (0); student is born in country of study but none of the parents are (1); student is born outside the country of study (2)
Economic, social and cultural status (ESCS)	Index based on answers to several questions on home possessions, parental occupation, parental education, aligned with the methodology to build the same index for PISA (Avvisati, 2020[70]; OECD, 2012[63])
Home location	Village (0), small town (1), town (2), city (3), large city (4)
English use in home	Binomial variable indicating whether students use English in their home
English exposure and use through traditional and new media	Index based on answers to several questions on frequency of use/exposure to English through traditional and new media

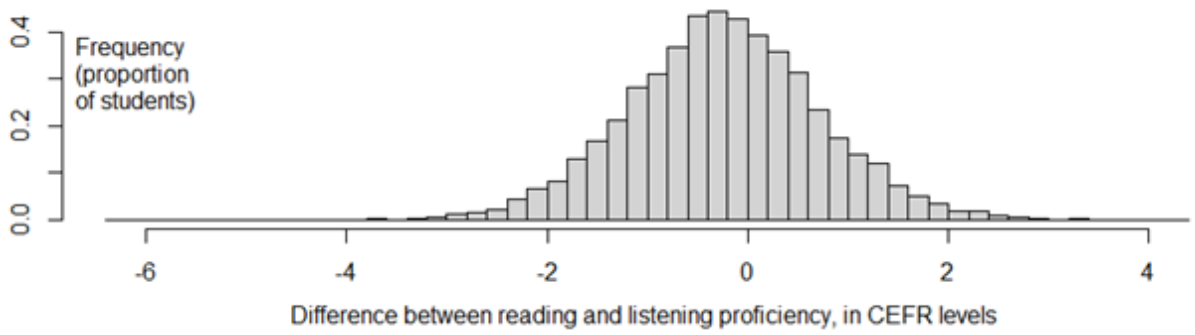
Note: As compared to the naming proposed by Jones et al. (2012[3]), "Target language" has been replaced with "English" to simplify the index names.

Finally, it is important to mention that the regression models include a binary variable for each education system, taking value 1 for students in that education system and 0 for other students (so-called "fixed effects"). This allows to control for any unobserved variable that is approximately constant within an education system. This include for example the grade of the student, since this was the same for all students in each education system. It also (approximately) include linguistic distance between English and the main language spoken by the student. This is approximately constant within education systems, because for the large majority of students, the language they speak best coincides with the instruction language of their education system.

Results

To put into perspective the regression results, it is useful to start by describing the distribution of the dependent variable (Figure 1). On average across the students in the sample, the difference between reading and listening skills is equal to –0.09 CEFR levels. This implies that the average student is slightly behind (by one tenth of a CEFR level) in English reading, compared to listening. This average difference is significant at the 1% level, even though the standard deviation (0.47) is relatively large. Overall, 62% of students in the sample demonstrate lower levels of reading than listening skills in English.

Figure 1. Difference between reading and listening skills among the students in the sample



The average difference between reading and listening skills is negative in almost all education systems. There is still a large variation across education systems, ranging from –0.67 in the Netherlands to 0.15 in Spain, where students are on average better at reading than at listening.

Table 6. Difference between reading and listening skills, by education system

	Mean	Standard deviation	Standard error
Bulgaria	-0.337	1.029	0.044
Croatia	-0.412	0.894	0.038
Estonia	-0.029	0.891	0.038
Flemish Community of Belgium	-0.362	0.916	0.039
France	-0.066	0.853	0.038
French Community of Belgium	0.058	0.824	0.037
German Community of Belgium	-0.096	0.954	0.062
Greece	-0.118	1.086	0.047
Malta	-0.387	1.09	0.055
Netherlands	-0.667	0.904	0.042
Poland	-0.132	0.924	0.038
Portugal	-0.363	0.897	0.039
Slovenia	-0.541	0.945	0.041
Spain	0.147	0.89	0.037
Sweden	-0.356	1.094	0.048

As for the empirical results from the baseline regression model, they support the main hypothesis (*H1a* and *H1b*) that learning English in the classroom is associated with relatively strong reading skills, compared to listening skills (Table 7). The reading-listening skill difference decreases on average by 0.0093 CEFR levels per year of English learning in formal education and by 0.00044 CEFR levels for a 1% increase in lesson time during the current year, after controlling for age, gender, socio-economic and immigration background, home location, use of English at home and through traditional and new media, and education system fixed effects (Column 4). These coefficients are not very large, but they are significant and not negligible. For example, a respondent that studied English for 3 years longer than a fellow student and spends 50% more time learning English in the classroom (e.g. 3 hours per week instead of 2) is expected to be ahead in her English reading skills by 0.046 CEFR levels more than in her English listening skills. This is about half as the average difference between reading and listening presented in Figure 1. Therefore, according to this model, European students would halve the gap between English reading and listening, if on average they had studied English for three additional years and they were enrolled in English lessons for 50% more time (in this hypothetical case, they would also be expected be at a higher level in both skills – see (European Commission, 2012[4])).

As expected (*H2a* and *H2b*), there is also a negative, significant association between the reading-listening skill difference and the number of trips abroad and face-to-face use of English (Column 3). These activities are thought to improve listening skills more than they improve reading skills, consistently with the evidence in Table 7.

Hypothesis *H3*, stating that more emphasis on teaching written and oral English is associated with, respectively, a larger and a smaller difference between reading and listening skills, is also supported by the evidence in Table 7 (Column 2). Both the coefficients for *emphasis on teaching oral English* and *emphasis on teaching written*

English are significant and with the expected sign. An increase by one standard deviation in the emphasis on written English and a decrease by one standard deviation in the emphasis on oral English would be associated with a reduction in the English reading-listening skill difference by 0.064, about two thirds of the average difference in the sample.

The complete baseline model is presented in Column 1 of Table 7. Besides the results already discussed, Column (1) shows that *H4* receives some support. The coefficient for *students' reported use of English during English lessons* is positive and significant at the 1% level. The coefficient for *emotional engagement* is also positive, but it is not significant at the 5% level. This could also be due to the collinearities induced by the inclusion of a large number of variables in the model. If *students' reported use of English during English lessons* is excluded from the baseline model, then the coefficient for *emotional engagement* becomes significant. This is consistent with the argument put forward in the research hypotheses section that these two variables could be interpreted as underlying indicators for the quality of the English lessons. Their joint significance in the model could indicate a positive association between quality of English formal instruction and the reading-listening skill difference.

Hypothesis *H5* is also supported by the data: *use of ICT in the classroom* is associated with a larger increase in listening than reading skills (i.e., it has a negative coefficient), while the contrary is true for *use of the textbook in the classroom*. A combination of an increase in the use of ICT and a decrease in the use of textbooks, both by one standard deviation, would be associated with a reduction by 0.048 CEFR levels in the reading-listening skill difference, about half the size of the skill gap observed in the sample. This result is particularly interesting, given that the data predates the Covid health crisis. The use of ICT has increased during the crisis (Marconi et al., 2020[37]), and it will be of great interest to test with future data whether this resulted in a reduction of the English reading-listening skill gap.

Among the control variables, only socio-economic background (ESCS) appears consistently significant across the various models (with a positive coefficient). Given that ESCS is associated with higher levels of all English skills in the dataset used in this paper (Azzolini, Campregher and Madia, 2020[71]), this means that on average students with higher ESCS do even better at reading than at listening, compared to other students. In contrast, the other control variables are not consistently significant. This is the case also for gender, even though it is generally found to be associated with foreign language skills (Azzolini, Campregher and Madia, 2020[71]; Denies et al., 2022[72]). It is interesting to note that both Azzolini et al. (2020[71]) and Denies et al. (2022[72]) find that gender is particularly associated with writing skills, and less so with reading and listening skills. This could explain the lack of significance for the variable gender in Table 7.

Table 7. Regression results

	(1) Dep. Var.: ExpDiff	(2) Dep. Var.: ExpDiff	(3) Dep. Var.: ExpDiff	(4) Dep. Var.: ExpDiff	(5) Dep. Var.: Reading	(6) Dep. Var.: Listening
Duration of English learning	0.0088 (0.0036)**	0.0093 (0.0036)**	0.0095 (0.0037)***	0.0093 (0.0037)**	0.0316 (0.0071)***	0.0157 (0.0055)***
English lesson time a week (log)	0.045 (0.020)**	0.049 (0.020)**	0.049 (0.020)**	0.044 (0.020)**	0.105 (0.071)	0.020 (0.066)
Emphasis on teaching oral English	-0.031 (0.012)***	-0.027 (0.011)**				
Emphasis on teaching written English	0.029 (0.014)**	0.037 (0.014)***				
Number of trips abroad	-0.014 (0.006)**	-0.016 (0.006)**	-0.016 (0.006)**			
Face-to-face use of English in home country	-0.026 (0.009)***	-0.028 (0.009)***	-0.029 (0.009)***			
Use of ICT in the classroom	-0.033 (0.007)***					
Use of the textbook in the classroom	0.012 (0.010)					
Students' reported use of English during English lessons	0.024 (0.008)***					
Emotional engagement	0.010 (0.008)					
Economic, social and cultural status (ESCS)	0.030 (0.010)***	0.034 (0.010)***	0.034 (0.010)***	0.026 (0.009)***	0.138 (0.013)***	0.107 (0.017)***
Home location	0.014 (0.011)	0.015 (0.011)	0.016 (0.011)	0.015 (0.011)	0.051 (0.018)***	0.034 (0.015)**
Immigration background	-0.017 (0.018)	-0.018 (0.018)	-0.018 (0.018)	-0.024 (0.018)	-0.046 (0.039)	-0.058 (0.036)
English use in home	0.011 (0.029)	0.016 (0.029)	0.017 (0.029)	0.0002 (0.028)	0.053 (0.048)	0.143 (0.041)***
Gender	-0.014 (0.017)	-0.021 (0.016)	-0.023 (0.016)	-0.021 (0.016)	-0.066 (0.038)*	0.010 (0.032)
Age	-0.018 (0.012)	-0.020 (0.012)	-0.020 (0.012)	-0.022 (0.012)*	-0.060 (0.026)**	-0.034 (0.025)
English exposure and use through traditional and new media	0.009 (0.012)	0.009 (0.012)	0.013 (0.012)	-0.002 (0.013)	0.165 (0.023)***	0.192 (0.024)***
Listening skills					0.658 (0.020)***	
Reading skills						0.533 (0.019)***
Inclusion of education system fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
N	6342	6342	6342	6342	6342	6342

Note: ***Significant at the 1% level; **Significant at the 5% level; *Significant at the 10% level. Standard errors account for clustering at the school level. ExpDiff is defined in Equation (2) as a measure of difference between reading and listening English skills.

The last two columns of Table 7 exemplify the methodological point made earlier on using plausible values both as dependent and independent variables. The models estimated are essentially the same as in Column (4). However, instead of estimating the association of the explanatory variables with the difference between reading and listening skills, they

estimate their association with reading conditional on listening skills (Column 5); and with listening conditional on reading skills (Column 6). Some straightforward algebraic manipulation[4] shows that, in absence of bias, the coefficient of every explanatory/ control variable in Column (5) should have the opposite sign as its coefficient in Column (6). However, this is not the case for most variables.

Based on the information reported in Columns (5) and (6), it is not possible to test directly the hypotheses put forward in this paper, which all focus on the difference in the association of some variable with reading and listening. To start with, it must be noted that the two models are not nested in each other, so (even abstracting from the bias) the two coefficients for a same explanatory variable cannot be compared to each other. In addition, it is not possible to test the hypothesis based only on one of the two models, because the conclusion would depend on the model used. For example, Column (5) reports a significant positive coefficient for Duration of English learning at school, which would seem to imply that this variable has a stronger association with reading than with listening. However, Column (6) reports a significant coefficient of the same sign, which would seem to imply the contrary (stronger association with listening than with reading).

Table 8 shows regression coefficients by education systems. As mentioned in the methodology section, the relatively small sample size does not allow a robust estimation of the model for each education system. This prevents a robust cross-country analysis on the role of linguistic distance between English and the main language spoken in each country, as done for example by Azzolini et al. (2020[71]). Nonetheless, it is instructive to look at cross-country variability, at least for the simplest model including Duration of English learning and English lesson time a week (in addition to control variables and education system fixed effects). The coefficients are positive for the large majority of education systems, both for Duration of English learning (12 out of 15) and for English lesson time a week (10 out of 15), even though they are rarely significant. There are, however, substantial differences across the coefficients. For example, the coefficient for English lesson time a week is negative and significant for Croatia. This implies a stronger association of lesson time with listening than reading skills, which is consistent with the strong emphasis on listening and speaking skills in this education system at the time of the survey (Brumen, Cagran and Rixon, 2009[16]).

[4] A multivariate regression model estimates an equation of this type:

Equation (Note3.1) $y_i = \beta_0 + \beta_1 \cdot x1_i + \beta_2 \cdot x2_i + \ldots + \beta_k \cdot xk_i + \varepsilon_i$

Where the subscript i=1,…,N denotes the unit of observation (the individual, in the case of this paper); yi is the dependent variable (in Column 5 of Table 7, this is reading); {x1i, …, xki} is the set of control and explanatory variables with associated coefficients {β1, …, βk}; and εi is a stochastic, individual-specific error term with mean 0. Now, suppose to run the same regression but using one of the control / explanatory variables (x1i) as the dependent variable and yi as a control variable (this is exactly what is done in Column 6 of Table 7, where listening is swapped with reading). Then the new estimation equation can be derived in terms of the parameters defined in Equation (Note3.1), and it is equal to:

Equation (Note3.2) $x1_i = -(\beta_0/\beta_1) + (1/\beta_1) \cdot y_i - (\beta_2/\beta_1) \cdot x2_i - \ldots - (\beta_k/\beta_1) \cdot xk_i - (1/\beta_1) \cdot \varepsilon_i$

Therefore, in absence of bias, the coefficients for the set of variables {x2i, …, xki} (i.e. those that always stay on the right-side of the equation) estimated from Equations (Note3.1) and (Note3.2) should always be of opposite sign.

Table 8. Regression coefficients for Duration of English learning and English lesson time a week, by education system (standard error in brackets)

	Duration of English learning	English lesson time a week (log)	Inclusion of control variables and education system fixed effects
Bulgaria	0.030 (0.009)***	0.087 (0.062)	Yes
Croatia	0.002 (0.013)	-0.235 (0.087)***	Yes
Estonia	0.006 (0.018)	0.096 (0.124)	Yes
Flemish Community of Belgium	0.030 (0.020)	-0.017 (0.142)	Yes
France	-0.010 (0.009)	0.075 (0.092)	Yes
French Community of Belgium	0.027 (0.013)**	0.319 (0.093)***	Yes
German Community of Belgium	-0.003 (0.027)	0.297 (0.11)***	Yes
Greece	-0.009 (0.015)	0.028 (0.062)	Yes
Malta	0.011 (0.013)	0.081 (0.142)	Yes
Netherlands	0.006 (0.015)	0.097 (0.118)	Yes
Poland	0.001 (0.015)	0.146 (0.110)	Yes
Portugal	0.031 (0.017)*	-0.0167 (0.067)	Yes
Slovenia	0.042 (0.024)*	-0.216 (0.150)	Yes
Spain	0.013 (0.012)	0.152 (0.089)*	Yes
Sweden	0.005 (0.012)	-0.110 (0.087)	Yes

Note: ***Significant at the 1% level; **Significant at the 5% level; *Significant at the 10% level. The model estimated for each education system is the same as the one reported for the full sample in Column (4) of Table 7.

Finally, before moving on with the conclusions, it is useful to note that the results presented in this section are robust to changes in the statistical methodology. In an online annex, the author calculated plausible values by bootstrapping across items and carried out the regression with a different weighting specification, obtaining very similar results (Marconi, 2022[73]).

Conclusions and implications

There seems to be a belief, among many linguists, that English learning in the classroom stimulates more the acquisition of reading skills, as compared to listening skills. The main goal of this paper was to derive some formal research hypotheses related to this belief, and test them statistically.

Five main hypotheses were put forward:

- More English learning in the classroom is associated with relatively high reading skills, compared to listening skills
- This is in contrast to other type of learning, like face-to-face interaction outside the classroom
- Teachers' emphasis on written or oral skills is also associated with the reading-listening skill difference
- Higher quality of instruction (or: better implementation of CLT), as proxied by using more English in the classroom and higher levels of student engagement, is associated with relatively high reading skills

- The use of different materials is also associated with the reading-listening skill difference (in particular, using ICT is associated with relatively high listening skills; and using the textbook with relatively high reading skills)

The hypotheses were tested on the European Survey of Language Competencies (ESLC, sometimes also known as Surveylang) dataset. This provides high-quality data on a randomised, representative sample of 14- and 15-year-old students in 16 European education systems. The fact that a subsample of students was tested on both reading and listening skills, and that all students answered to an extensive background questionnaire, made it possible to build all the measures needed for this study.

All the five hypotheses were supported by the data, even after controlling for a wide range of student characteristics and for education system fixed effects. This provides robust evidence for a stronger association of formal English instruction with reading, rather than listening skills. In particular: given a certain level of listening skills, the more years students have been learning English at school, the better their reading skills; the more hours of English lesson students are currently attending, the better their reading skills; and the better the lessons, the better their reading skills, again in comparison to their listening skills.

Some methodological and policy implications can be drawn from the results presented in this paper. From a methodological point of view, these results suggest caution when using plausible values as control variables. The existing literature on this subject did not reach a unanimous conclusion, with some authors warning about the bias introduced when doing so (Junker, Schofield and Taylor, 2012[57]; Schofield et al., 2014[58]); and others claiming that the bias is not necessarily very large (Braun and von Davier, 2017[53]). The estimates presented in this paper suggest that the bias introduced by controlling for a variable measured with error is large, in the context of the data and research design employed in this study.

In terms of policy, this paper contributes system-level evidence showing that large-scale changes in teaching practices are associated with changes in specific English skills (reading and listening) learned by students. This strengthens the existing evidence that there is scope for policy to intervene and influence learning at the level of the specific English skill. This is not something to be taken for granted. There is a lot of evidence in the literature from experiments on methodologies to teach specific skills, but this literature usually focus on small-scale and convenience samples (Andringa and Godfroid, 2020[26]). The social sciences are often faced with the problem of bridging the evidence on phenomena happening at the micro- and macro-level (Bouvier, 2011[74]; Aguinis et al., 2011[75]). This makes system-level evidence, collected from representative large-scale samples, necessary in the process of knowledge building (Cumming, 1996[76]). This is the main rationale for the undertaking of the PISA 2025 Foreign Language Assessment (OECD, 2021[1]), which will have the crucial advantage of covering countries beyond Europe, all around the world.

While maintaining overall policies that considers all communicative skills (listening, reading, speaking and writing) important in the teaching of foreign languages, governments already have some differentiation in place (Eurydice, 2017[65]). For example, oral skills are given more emphasis in the first phases of foreign language learning in a few countries, while writing receives less weight than other skills in Switzerland. The choice of the skills to be assessed in the OECD's PISA Foreign Language Assessment (OECD, 2021[1]) has been made based on countries' policy priorities, another sign that governments have their own priorities over the communicative skills to be learned in schools. The results of this paper on the association of generic schooling, teaching materials and teaching emphasis with the learning of reading and listening add to the evidence on the education policy tools that can be used to put this priorities into practice.

Finally, it is important to note that the results presented in this paper refer to specific countries and time. The countries participating to the ESLC survey were all European. Equally importantly, much has changed across education systems in the last decade, with the Covid crisis both forcing and acting as a catalyst for change (Marshall, Pressley and Love, 2022[27]; Schleicher, 2022[28]). In the future, it would be interesting to investigate the hypotheses put forward in this study for different countries and a more recent time period. The new data provided by the PISA 2025 Foreign Language Assessment (OECD, 2021[1]) will give a chance to address this limitation.

References

Aguinis, H. et al. (eds.) (2011), "Walking New Avenues in Management Research Methods and Theories: Bridging Micro and Macro Domains", *Journal of Management*, Vol. 37/2, pp. 395-403, https://doi.org/10.1177/0149206310382456. [75]

Ahmed, S. (2017), *Authentic ELT Materials in the Language Classroom: An Overview*. [45]

Amengual-Pizarro, M. (2009), "Does the English Test in the Spanish University Entrance Examination Influence the Teaching of English?", *English Studies*, Vol. 90/5, pp. 582-598, https://doi.org/10.1080/00138380903181031. [13]

Andringa, S. and A. Godfroid (2020), "Sampling Bias and the Problem of Generalizability in Applied Linguistics", *Annual Review of Applied Linguistics*, Vol. 40, pp. 134-142, https://doi.org/10.1017/s0267190520000033. [26]

Avvisati, F. (2020), "The measure of socio-economic status in PISA: a review and some suggested improvements", *Large-scale Assessments in Education*, Vol. 8/1, https://doi.org/10.1186/s40536-020-00086-x. [70]

Azzolini, D., S. Campregher and J. Madia (2020), "Formal instruction vs informal exposure. What matters more for teenagers' acquisition of English as a second language?", *Research Papers in Education*, Vol. 37/2, pp. 153-181, https://doi.org/10.1080/02671522.2020.1789718. [71]

Berkleyen, N. (2007), "An investigation of English teacher candidates' problems related to listening.", *Electronic Journal of Social Sciences*, Vol. 6, pp. 91-105, https://dergipark.org.tr/en/download/article-file/69971. [12]

Bouvier, A. (2011), "Individualism, Collective Agency and the "Micro–Macro Relation"", in *The SAGE Handbook of the Philosophy of Social Sciences*, SAGE Publications Ltd, 1 Oliver's Yard, 55 City Road, London EC1Y 1SP United Kingdom , https://doi.org/10.4135/9781473913868.n9. [74]

Bozorgian, H. (2012), "The Relationship between Listening and Other Language Skills in International English Language Testing System", *Theory and Practice in Language Studies*, Vol. 2/4, https://doi.org/10.4304/tpls.2.4.657-663. [7]

Braun, H. and M. von Davier (2017), "The use of test scores from large-scale assessment surveys: psychometric and statistical considerations", *Large-scale Assessments in Education*, Vol. 5/1, https://doi.org/10.1186/s40536-017-0050-x. [53]

Brown, J. (1954), "How teachable is listening?", *Educational Research Bulletin*, Vol. 33, pp. 85-93, http://www.jstor.org/stable/1473241. [34]

Brumen, M., B. Cagran and S. Rixon (2009), "Comparative assessment of young learners' foreign language competence in three Eastern European countries", *Educational Studies*, Vol. 35/3, pp. 269-295, https://doi.org/10.1080/03055690802648531. [16]

Butler, Y. (2004), "What Level of English Proficiency Do Elementary School Teachers Need to Attain to Teach EFL? Case Studies from Korea, Taiwan, and Japan", *TESOL Quarterly*, Vol. 38/2, p. 245, https://doi.org/10.2307/3588380. [41]

Buttler, Y. (2007), "How Are Nonnative-English-Speaking Teachers Perceived by Young Learners?", *TESOL Quarterly*, Vol. 41/4, pp. 731-755, https://doi.org/10.1002/j.1545-7249.2007.tb00101.x. [9]

Celce-Murcia, M. (2014), "An overview of language teaching methods and approaches", in *Teaching English as a second or foreign language*. [36]

Chacón, C. (2005), "Teachers' perceived efficacy among English as a foreign language teachers in middle schools in Venezuela", *Teaching and Teacher Education*, Vol. 21/3, pp. 257-272, https://doi.org/10.1016/j.tate.2005.01.001. [44]

Chambers, F. (1991), "Promoting use of the target language in the classroom", *Language Learning Journal*, Vol. 4/1, pp. 27-31, https://doi.org/10.1080/09571739185200411. [50]

Cheng, L. (1997), "How Does Washback Influence Teaching? Implications for Hong Kong", *Language and Education*, Vol. 11/1, pp. 38-54, https://doi.org/10.1080/09500789708666717. [14]

Choi, E. and J. Lee (2016), "Investigating the relationship of target language proficiency and self-efficacy among nonnative EFL teachers", *System*, Vol. 58, pp. 49-63, https://doi.org/10.1016/j.system.2016.02.010. [43]

Council of Europe (2018), *Common European framework of reference for languages: Learning, teaching, assessment – Companion volume with new descriptors.*, http://www.coe.int/lang-cefr. [23]

Council of Europe (2001), *Common European framework of reference for languages: Learning, teaching, assessment.*, https://rm.coe.int/16802fc1bf. [22]

Cumming, A. (1996), "IEA's Studies of Language Education: their scope and contributions", *Assessment in Education: Principles, Policy & Practice*, Vol. 3/2, pp. 179-192, https://doi.org/10.1080/0969594960030205. [76]

de Waal, T., J. Pannekoek and S. Scholtus (2011), *Handbook of Statistical Data Editing and Imputation*, Wiley, https://doi.org/10.1002/9780470904848. [69]

Denies, K. et al. (2022), "Mapping and explaining the gender gap in students' second language proficiency across skills, countries and languages", *Learning and Instruction*, Vol. 80, p. 101618, https://doi.org/10.1016/j.learninstruc.2022.101618. [72]

Dolan, E. (ed.) (2016), "Rasch Analysis for Instrument Development: Why, When, and How?", *CBE—Life Sciences Education*, Vol. 15/4, p. rm4, https://doi.org/10.1187/cbe.16-04-0148. [66]

European Commission (2012), *First european survey on language competences – Final Results.*, https://op.europa.eu/en/publication-detail/-/publication/42ea89dc-373a-4d4f-aa27-9903852cd2e4/language-en/format-PDF/source-116835286. [4]

Eurydice (2017), *Key data on teaching languages at school in Europe – 2017.*, https://doi.org/10.2797/839825. [65]

Eurydice and Eurostat (2012), *Key data on teaching languages at school in Europe – 2012.*, https://doi.org/10.2797/12090. [64]

Funk, H. (2012), "Four models of language learning and acquisition and their methodological implications for textbook design", *Electronic Journal of Foreign Language Teaching*, Vol. 9/SUPPL.1. [47]

Gilakjani, A. and N. Sabouri (2016), "Learners' Listening Comprehension Difficulties in English Language Learning: A Literature Review", *English Language Teaching*, Vol. 9/6, p. 123, https://doi.org/10.5539/elt.v9n6p123. [11]

Gilmore, A. (2007), "Authentic materials and authenticity in foreign language learning", *Language Teaching*, Vol. 40/2, pp. 97-118, https://doi.org/10.1017/s0261444807004144. [46]

Goh, C. (2008), "Metacognitive Instruction for Second Language Listening Development", *RELC Journal*, Vol. 39/2, pp. 188-213, https://doi.org/10.1177/0033688208092184. [29]

Goldspink, C. and M. Foster (2013), "A conceptual model and set of instruments for measuring student engagement in learning", *Cambridge Journal of Education*, Vol. 43/3, pp. 291-311, https://doi.org/10.1080/0305764x.2013.776513. [51]

Graves, K. and S. Garton (2017), "An analysis of three curriculum approaches to teaching English in public-sector schools", *Language Teaching*, Vol. 50/4, pp. 441-482, https://doi.org/10.1017/s0261444817000155. [38]

Green, A. and M. Bracciodieta (2019), *Interactive listening – A can-do Paradigm.*, https://www.bib.irb.hr/1064953/download/1064953.11_Kovai_Buba_2019_Relationships_between_Students_LANAK_U_ZBORNIKU.pdf#page=470. [8]

Greene, W. (2003), *Econometric Analysis*, Prentice Hall. [20]

Hino, N. (1988), "Yakudoku: Japan's dominant tradition in foreign language learning", *JALT Journal*, Vol. 10/1. [31]

Hiver, P., A. Al-Hoorie and S. Mercer (eds.) (2021), *Student Engagement in the Language Classroom*, Multilingual Matters, https://doi.org/10.21832/9781788923613. [49]

Hosoki, Y. and H. Yukiko (2011), "English Language Education in Japan : Transitions and Challenges (I)", *紀要論文(ELS) / Departmental Bulletin Paper*, Vol. 6. [32]

Isaacs, T. (2012), "Teaching and Learning Second Language Listening: Metacognition in Action (review)", *The Canadian Modern Language Review / La revue canadienne des langues vivantes*, Vol. 68/3, pp. 349-351, https://muse.jhu.edu/article/483689/pdf#info_wrap. [10]

Jolliffe, I. (2002), *Principal component analysis. Springer Series in Statistics.* [68]

Jones, N. et al. (2012), *First European Survey on Language Competences – Technical Report.*, https://ec.europa.eu/assets/eac/languages/policy/strategic-framework/documents/language-survey-technical-report_en.pdf. [3]

Junker, B., L. Schofield and L. Taylor (2012), "The use of cognitive ability measures as explanatory variables in regression analysis", *IZA Journal of Labor Economics*, Vol. 1/1, https://doi.org/10.1186/2193-8997-1-4. [57]

Klepper, S. (1988), "Regressor diagnostics for the classical errors-in-variables model", *Journal of Econometrics*, Vol. 37/2, pp. 225-250, https://doi.org/10.1016/0304-4076(88)90004-8. [19]

Klepper, S., M. Kamlet and R. Frank (1993), "Regressor Diagnostics for the Errors-in-Variables Model - An Application to the Health Effects of Pollution", *Journal of Environmental Economics and Management*, Vol. 24/3, pp. 190-211, https://doi.org/10.1006/jeem.1993.1013. [59]

Li, D. (1998), ""It's Always More Difficult Than You Plan and Imagine": Teachers' Perceived Difficulties in Introducing the Communicative Approach in South Korea", *TESOL Quarterly*, Vol. 32/4, p. 677, https://doi.org/10.2307/3588000. [40]

Linacre, J. (1999), "Understanding Rasch measurement: estimation methods for Rasch measures.", *Journal of outcome measurement*, Vol. 3/4. [62]

Lumley, T. (2021), "Analysis of Complex Survey Samples.", *CRAN Package Repository*, https://cran.r-project.org/web/packages/survey/survey.pdf. [55]

Lumley, T. (2004), "Analysis of Complex Survey Samples", *Journal of Statistical Software*, Vol. 9/8, https://doi.org/10.18637/jss.v009.i08. [54]

Marconi, G. (2022), "Does English Instruction Teach More Reading Than Listening Skills? Evidence From 15 European Education Systems", *SSRN Electronic Journal*, https://doi.org/10.2139/ssrn.3998936. [73]

Marconi, G. (2018), "Education as a Long-Term Investment: The Decisive Role of Age in the Education-Growth Relationship", *Kyklos*, Vol. 71/1, pp. 132-161, https://doi.org/10.1111/kykl.12165. [60]

Marconi, G. et al. (2020), "What matters for language learning?: The questionnaire framework for the PISA 2025 Foreign Language Assessment", *OECD Education Working Papers*, No. 234, OECD Publishing, Paris, https://doi.org/10.1787/5e06e820-en. [37]

Marconi, G., L. Vergolini and F. Borgonovi (2023), "The demand for language skills in the European labour market: Evidence from online job vacancies", *OECD Social, Employment and Migration Working Papers*, No. 294, OECD Publishing, Paris, https://doi.org/10.1787/e1a5abe0-en. [2]

Marshall, D., T. Pressley and S. Love (2022), "The times they are a-changin':Teaching and learning beyond COVID-19", *Journal of Educational Change*, Vol. 23/4, pp. 549-557, https://doi.org/10.1007/s10833-022-09469-z. [27]

Miller, L. (2001), "Developing listening skills with authentic materials.", *ESL Magazine*, p. 689, http://dl.ueb.edu.vn/bitstream/1247/9968/1/Developing%20Listening%20Skills%20with%20Authentic%20Materials%281%29.pdf. [15]

Ngoc Khoi, M. (2014), "Towards a holistic approach to developing the language proficiency of Vietnamese primary teachers of English", *Electronic Journal of Foreign Language Teaching*, Vol. 11/2. [42]

OECD (2021), *PISA 2018 technical report.*, https://www.oecd.org/pisa/data/pisa2018technicalreport/. [24]

OECD (2021), *PISA 2025 Foreign Language Assessment Framework*, https://www.oecd.org/pisa/foreign-language/PISA-2025-FLA-Framework.pdf. [1]

OECD (2019), *PISA 2018 Results (Volume I): What Students Know and Can Do*, PISA, OECD Publishing, Paris, https://doi.org/10.1787/5f07c754-en. [21]

OECD (2012), *PISA 2009 Technical Report*, PISA, OECD Publishing, Paris, https://doi.org/10.1787/9789264167872-en. [63]

OECD (2009), *PISA Data Analysis Manual: SPSS, Second Edition*, PISA, OECD Publishing, Paris, https://doi.org/10.1787/9789264056275-en. [52]

Oxford, R. (1993), "Research update on teaching L2 listening", *System*, Vol. 21/2, pp. 205-211, https://doi.org/10.1016/0346-251x(93)90042-f. [5]

Pan, T. (2018), "Fitting the Rasch model under the logistic regression framework to reduce estimation bias", *Journal of Modern Applied Statistical Methods*, Vol. 17/1, https://doi.org/10.22237/jmasm/1530028025. [61]

Philp, J. and S. Duchesne (2016), "Exploring Engagement in Tasks in the Language Classroom", *Annual Review of Applied Linguistics*, Vol. 36, pp. 50-72, https://doi.org/10.1017/s0267190515000094. [48]

Richards, J. and T. Rodgers (2001), *Approaches and Methods in Language Teaching*, Cambridge University Press, https://doi.org/10.1017/cbo9780511667305. [33]

Savignon, S. (2017), *Communicative Competence*, John Wiley & Sons, Inc., Hoboken, NJ, USA, https://doi.org/10.1002/9781118784235.eelt0047. [39]

Schleicher, A. (2022), *Building on COVID-19's Innovation Momentum for Digital, Inclusive Education*, International Summit on the Teaching Profession, OECD Publishing, Paris, https://doi.org/10.1787/24202496-en. [28]

Schofield, L. et al. (2015), "Predictive Inference Using Latent Variables with Covariates", *Psychometrika*, Vol. 80/3, https://doi.org/10.1007/s11336-014-9415-z. [18]

Schofield, L. et al. (2014), "Predictive Inference Using Latent Variables with Covariates", *Psychometrika*, Vol. 80/3, pp. 727-747, https://doi.org/10.1007/s11336-014-9415-z. [58]

Sok, S., H. Shin and J. Do (2021), "Exploring which test-taker characteristics predict young L2 learners' performance on listening and reading comprehension tests", *Language Testing*, Vol. 38/3, pp. 378-400, https://doi.org/10.1177/0265532221991134. [56]

Spoden, C., J. Fleischer and M. Leucht (2020), "Converging Development of English as Foreign Language Listening and Reading Comprehension Skills in German Upper Secondary Schools", *Frontiers in Psychology*, Vol. 11, https://doi.org/10.3389/fpsyg.2020.01116. [17]

Sudina, E. (2021), "Study and Scale Quality in Second Language Survey Research, 2009–2019: The Case of Anxiety and Motivation", *Language Learning*, Vol. 71/4, pp. 1149-1193, https://doi.org/10.1111/lang.12468. [25]

Taber, K. (2017), "The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education", *Research in Science Education*, Vol. 48/6, pp. 1273-1296, https://doi.org/10.1007/s11165-016-9602-2. [67]

Tschirner, E. (2016), "Listening and Reading Proficiency Levels of College Students", *Foreign Language Annals*, Vol. 49/2, pp. 201-223, https://doi.org/10.1111/flan.12198. [6]

Whitley, M. (1993), "Communicative Language Teaching: An Incomplete Revolution", *Foreign Language Annals*, Vol. 26/2, pp. 137-154, https://doi.org/10.1111/j.1944-9720.1993.tb01162.x. [35]

Zhou, S., P. Hiver and A. Al-Hoorie (2021), "Measuring L2 Engagement: A Review of Issues and Applications", in *Student Engagement in the Language Classroom*. [30]