The meaningfulness of two curriculum-based national tests of English as a foreign language

Raili Hilden
Department of Education, University of Helsinki, Finland

Jonna Pulkkinen
Finnish Institute for Educational Research, University of Jyväskylä, Finland

Juhani Rautopuro
Finnish Institute for Educational Research, University of Jyväskylä, Finland

This paper addresses the aspect of the meaningfulness of a national assessment in English as a foreign language, applying the fairness framework proposed by Kunnan (2018). We compared students' performance on two receptive language skills of listening and reading on two subsequent national evaluations in Finland, taken by students at the end of compulsory basic education and at the end of general upper secondary education, respectively. The research questions focus on (1) the relationship between students' performance on the two tests and their gender, language of schooling and parents' educational level, and (2) the relationship between the students' receptive language proficiency at the end of basic education and general upper secondary education. The data were analysed using linear regression and quantile regression analyses. The effect of background variables on the proficiency was stronger for low-performing students than for high-performing students. Moreover, students' proficiency on the receptive skills at the end of basic education well predicted that at the end of general upper secondary education across several points in the score distribution. The findings also indicated persistent challenges with respect to educational equality and equity that requires ongoing attention by policy-makers and test designers.

Email address for correspondence: raili.hilden@helsinki.fi

Key words: large-scale assessment, learning outcomes, listening comprehension, reading comprehension, validation, fairness, justice

Introduction

The goal of this paper is twofold: first, we investigated the equality and equity of learning outcomes in English as a foreign language, and secondly, we built a validity argument for the national assessment based on two subsequent evaluations of one student cohort. The first dataset was drawn from national assessments at the end of compulsory basic education at the end of year 9, and the second one from a school leaving examination at the end of general upper secondary education, the Finnish Matriculation exam (ME). The two assessments represented two kinds of summative assessment, one low-stakes and the other high-stakes. Both of them claim to measure the construct of communicative language proficiency, based on the Common European Framework of Reference (CEFR)-adapted scale ranging from Level A1.1 to C1.1 (Council of Europe, 2001; Hildén & Takala, 2007).

In Finland, national evaluations of learning outcomes (NELO) are mandated by the Finnish Ministry of Education and Culture and administered by the Finnish Education Evaluation Centre in different school subjects once in a curriculum period of a decade. The aim of these evaluations is to measure the extent to which the language proficiency goals, set in the operative curriculum are attained by a cohort of school leavers at the end of basic education. These program evaluations are carried out to assure the quality of basic education, but the outcomes are used merely for guidance and informative purposes. Feedback is provided to schools to indicate their relative position among all schools in the sample, but the results have no substantial impact for either teachers or their pupils.

A salient feature of the Finnish national assessments that is worth mentioning is their aim to monitor the implementations of equity and equality in education (Pizorn & Huhta 2016, p. 244). In Finland, the equity of education is defined as a socio-political ideal guiding the availability of schooling as well as its outcomes of equipping all school leavers with sufficient competencies to enter society irrespective of their gender, social background or the language of schooling. The

equality law aims at removing all kinds of gender-related discrimination and improving women´s rights, particularly in workforce life. The non-discrimination act seeks to prevent discrimination on any grounds. Both laws obligate educational organizations to make and update a plan for implementing equity and non-discrimination (Finnish National Agency of Education, 2022).

A large-scale evaluation of modern foreign languages was conducted in 2013 comprising a total of eight syllabi, based on the core curriculum in effect since 2004 (Finnish National Board of Education, 2004). In the Finnish educational system, many foreign languages are offered to pupils, but in practice, the most commonly studied foreign language is English. In the Finnish education system, all pupils continue studying either in general upper secondary education, which is an academic track, or in vocational upper secondary education after completing compulsory basic education. About half of the pupils from basic education move to general upper secondary education (Statistics Finland, 2020), and they are the group addressed in this study.

In general, studies in the general upper secondary education include around 75 courses with 38 study hours for each course. The aims of this educational stage comprise general objectives, such as enabling the students to grow into an educated member of society, as well as allowing them to acquire subject-specific knowledge and skills (Finnish National Board of Education, 2003; 2015). The English advanced syllabus comprises six mandatory courses, but most students choose to take one or two optional courses on the top of these.

At the end of general upper secondary education, students take the Matriculation Exam (ME), the only high-stakes test administered on a large scale in Finland. The ME is not compulsory in principle, but in practice all students completing general upper secondary education take this exam (about 30,000 every year). Throughout its 170 years of history, the ME is a flagship of Finnish education and an important landmark in adolescents' private lives. The ME is run twice a year and covers more than forty tests in multiple syllabi in subjects taught in general upper secondary schools (Finnish Matriculation Examination Board, 2020). In contrast to NELO, test-takers' performance on the ME bears substantial consequences for their future lives and for admissions to tertiary education

programmes. Therefore, it is vitally important to examine the fairness and justice of this exam (Kunnan, 2018).

Literature review

Modern validity theory

The domain of consequential validity, originally introduced by Messick (1989), has inspired an ever-strengthening discourse voicing the ancient principles of fairness and justice as ultimate determinants of all evaluations. Kunnan (2018) states that fairness is the prior condition of social justice in test administration and use. Fairness builds on four primary sub-principles: (a) adequate opportunities, (b) context and interpretation, (c) absence of bias, and (d) appropriate access, administration and standard setting procedures. These four principles, again, contribute to justice at social level. The two sub-principles of justice incorporate social benefit and positive impact on the one hand, and positive values to enhance justice in society on the other. Building an argument (Toulmin, 1958; Kane, 2006, 2012; Bachman & Palmer, 2010) to support the meaningfulness claim starts with the principle of fairness that frames the entire procedure of a test cycle from test development to the consequences of test use.

This study focuses on the aspect of meaningfulness that can be explored by gathering evidence from "the content, criterion, construct and consequences of an assessment" (Kunnan, 2018, pp. 139-140). In fact, Kunnan uses meaningfulness as a term that is synonymous with validity, but it is contested to alter and replace the well-established term validity and to reduce its conceptual content into meaningfulness (see e.g., Weideman, 2019). Although validity may not be entirely captured by meaningfulness, we found that Kunnan's rationale was suitable for our research aim, because it enables us to address the core principles of Finnish educational pursuits, i.e., equality and equity, in a transparent and tangible manner. Investigating meaningfulness in terms of reliability and consistency across tests measuring the same construct of communicative language ability (Kunnan, 2018, p. 96) lends insight into how these domains of fairness are attended to and how they contribute to social justice.

The core principle of fairness states that both the NELO and the ME should be fair to all test-takers by offering them equal opportunity to demonstrate their language proficiency. NELO and ME should be meaningful and consistent regarding test score interpretation for all sample pupils in terms of (a) curriculum objectives, (b) the construct of language ability, (c) language, content and topics, (d) being able to predict performance with respect to external criteria, (e) consistency within sets of items/tasks in terms of different constructs, (f) consistency across multiple assessment tasks, forms and/or occasions of assessments (in different regions, offices, and rooms), and (g) consistency across multiple examiners/raters (Kunnan, 2018, pp. 96-97). The claim is supported by warrants, whereas any contrasting findings are regarded as rebuttals. Both warrants and rebuttals are backed with empirical data.

In this study, the claim can be stated as following:

The NELO is meaningful and consistent in being able to predict performance in a subsequent high-stakes test in terms of the linguistic skills measured across gender, language of schooling and parental educational level.

Determinants of educational outcomes

Despite some research on the concurrent validity between curriculum-based test outcomes and school grades (Ouakrim-Soivio et al., 2018), the predictive accuracy of curriculum-based tests in foreign languages at the end of compulsory education has not often been reported in scholarly publications. The Nordic countries have similar systems of regular evaluations of learning outcomes in school subjects, with a few studies from Sweden and Denmark. Studies on national tests in Denmark reveal a strong relationship between students' national test results and their later educational outcomes. Furthermore, socio-economic gaps in achievement between children have been documented across all grade levels (Beuchert & Nandrup, 2017). Similar indications of inequalities have been suggested by Finnish and Swedish researchers (Korp, 2006). Differences in learning outcomes in favour of male students have been reported by (Börjesson & Nilsson, 2018), and in Finland better learning outcomes are achieved by

children of higher-educated families (Härmälä, Huhtanen & Puukko, 2014; Hildén & Rautopuro, 2014a; 2014b).

The predictive power of the Programme for International Student Assessment (PISA) for subsequent achievement in the subjects measured by PISA tests is steadily increasing, but to date, foreign languages have not been included in these international surveys. Yet the age of taking PISA evaluations coincide with the national evaluations, and therefore it is worth taking note of the major findings attesting the relatively strong predictive power of the PISA tests in terms of academic achievement and its indicators, such as school grades (Fischbach et al., 2013; Pulkkinen & Rautopuro, 2022).

In light of our research aim, a number of PISA studies have addressed gender differences, revealing substantial gaps between genders in terms of skills, grades and aspirations (e.g., Matějů & Smith, 2015). Performance at the end of compulsory basic education is associated with schooling and social factors, which, in turn, affects access to higher education (Murdoch et al., 2011). Furthermore, performance on the PISA tests significantly predicts literacy skills in adulthood ten years later (Rosdahl, 2014).

Language and literacy skills in large-scale assessments

Finnish PISA results in 2015 exhibited an association between L1 reading skills and the following background variables: reading engagement, reading enjoyment time, gender and socio-economic background. Girls tend to outperform boys in reading, and they also feature as more engaged readers of a variety of text types. Also, socio-economic background affects reading results significantly, and the association is both direct and mediated by an array of variables, such as cultural capital and positive attitudes towards reading (Leino & Nissinen, 2018). The cohort assessed in the PISA measurement round 2015 is close to the cohort assessed in this study. The most recent PISA studies show a slight decrease in overall reading skills among Finnish youngsters (Leino et al., 2021).

The earlier NELOs regularly addressed gender, language of instruction and parents' educational level by comparing the average performance between groups. The overall primacy of girls was prominent in advanced syllabi of Swedish

in Finnish language instruction schools (Hilden & Rautopuro, 2014a) and Finnish in Swedish language instruction schools (Toropainen, 2010) as well as in foreign languages such as German (Hilden & Rautopuro, 2014b) and French (Härmälä & Huhtanen, 2014).

Language of schooling is a relevant factor in the Finnish educational system because Finland is an officially bilingual country with Finnish and Swedish as the two domestic languages. The superiority of Swedish language schools is prominent in most studies of learning outcomes in all the frequently taught languages (Härmälä, Huhtanen, Silverström et al., 2014; Härmälä et al., 2019: 74 – 79). Also in the European Survey of Language Competences (European Commission, 2012), pupils in Sweden appeared to possess the highest levels of English proficiency (including in listening and reading proficiency addressed in our study) among all the participating countries. In contrast with the Finnish results in national evaluations of learning outcomes in English, no substantial differences exist between girls' and boys' achievement at the end of the Swedish basic education. However, towards the end of upper secondary education, boys outperform girls, even in Sweden, particularly on short-answer items or multiple-choice questions in listening and reading tests (Börjesson & Nilsson, 2018).

Differences based on parents' educational level (i.e., whether they have completed ME) exhibited a fairly conventional pattern, suggesting that children of higher educated parents scored higher than their peers in English (Härmälä, Huhtanen & Puukko, 2014; Härmälä et al., 2019: 74 – 79). The same pattern applies to Swedish tested in Finnish schools, as well as German (Hilden & Rautopuro, 2014a; 2014b) and French (Härmälä & Huhtanen, 2014).

This paper focuses on receptive language proficiency comprising listening and reading skills that are expected to develop from the average CEFR level B1.1 at the end of basic education to the level of B2.1 at the end of upper secondary education. Language users at Level B1 in listening are expected to understand straightforward factual information in everyday situations and about familiar topics, when the speech is clearly articulated and the accent in general is familiar. At Level B2, they can cope with more complex speech on various tangible and abstract topics, given the explicitness of argumentation in a standard dialect

(Council of Europe, 2001: 66). When it comes to reading, those at Level B1 can demonstrate a satisfactory level of comprehension of straightforward factual texts on familiar topics. At Level B2, they have a broad vocabulary, which enables them to read selectively and with a large degree of independence. They can vary their reading speed and style according to the purpose of the reading task (Council of Europe, 2001, p. 69). The CEFR levels and descriptors are indicated in the Finnish core curricula based on the version adapted and validated for basic and upper secondary education (Hildén & Takala, 2007).

The present study

The purpose of this study was to investigate the meaningfulness of national assessments of English as a foreign language in the Finnish educational context, thus laying the foundation for its fairness and justice. In this study, the NELO should be meaningful and consistent in being able to predict students' performance in a subsequent high-stakes test in terms of the linguistic skills measured across genders, languages of schooling and parental educational levels.

According to the traditional view on validity, our study addressed the reliability and predictive validity of national evaluations at the end of compulsory basic education in relation to a high-stakes exam administered to students when they finish the general upper secondary stage (Fulcher & Davidson, 2007, pp. 4–5). Students' scores on the ME serve as a criterion against which the corresponding language skills displayed in the national evaluation can be compared. By the same token, the NELO scores contribute to determining the meaningfulness of the ME, by setting a baseline of between-group comparisons.

If the relationship between the background variables and language proficiency detected at the end of basic education is maintained after upper secondary education, it suggests the predictive value of the NELO in relation to the ME. If the relationship remains or changes, the analysis sheds light on the added value of the general upper secondary language education regarding the amount of language proficiency and its power to even out, maintain or even increase groupwise differences. This information becomes an indication of the fairness of the two tests and also the educational path in between.

Among the multiple facets proposed by Kunnan (2018), this study addresses the fairness sub-claim of meaningfulness that is characterised as the ability of a test "to predict performance in terms of external criteria" (Kunnan, 2018, p. 96). We positioned the ME as an external criterion in relation to the NELO, and set out to explore whether test-takers' scores on the NELO are able to predict their performance on the ME, and whether such prediction is affected by variables including their gender, parents' educational level, and language of schooling. These background variables have been acknowledged as pivotal indicators of equality and equity in previous research and the Finnish legislation (Equality act, 2014; Equity act, 2014).

In this study, we focused on the following research questions:

RQ1. What is the relationship between the receptive language proficiency and certain background variables (gender, language of instruction and parents' educational level) in the NELO and ME?

RQ2. What is the relationship between the language proficiency in the NELO (at the end of basic education) and the language proficiency in the ME (at the end of general upper secondary education)?

The first research question scrutinises the claim considering the equality of performance across the three background variables, i.e., girls vs. boys, higher vs. lower educated parents, and Finnish vs. Swedish language instruction schools. This question does not directly address the validity or meaningfulness of the tests themselves, but rather the capacity of the preceding level of schooling to produce equal learning outcomes.

The claim to be probed is: *Compulsory basic education provides equal learning outcomes irrespective of gender, parental education level or language of instruction.*

The second research question concerns the relationship between two national assessments of English. The answers to this research question lend direct insight into the predictive power of the NELO. Furthermore, the comparison enlightens us of the impact of upper secondary language education. Formulating a claim for

RQ2 is not as straightforward as that for RQ1, since we do not know whether the relationships between the language proficiency and background variables at the end of basic education prevail after the upper secondary level, or whether they disappear or are reshaped. From the consistency perspective, maintenance of the relationship between background variables and language proficiency in focus would warrant the consistency claim presuming that the students' rank order would remain the same throughout upper secondary education. On the other hand, advocating more equal and equitable learning results, our best expectation would be that the ME results are more equal and group-wise differences smaller compared to the division of success in the NELO.

Materials and methods

Data

The data set comprised student performance data from national evaluations in English syllabi starting at the age of nine. The study is a longitudinal comparison of students' (N = 1,485) language ability at the end of compulsory basic education and in the general upper secondary school leaving exam. The data include students who participated in the national assessment of English proficiency at the end of basic education (NELO) in 2013 and in the Matriculation Exam (ME) at the end of general upper secondary education in 2015 or in the spring of 2016. Table 1 summarises the participants' background.

The data at the end of basic education were collected by using two-stage stratified random sampling. In the first phase, the schools were stratified based on region, municipality type and school size. Second, a random sample of students was drawn from within the schools in the sample. The final data consist of 3,476 students: 2,966 from 94 Finnish-speaking schools and 510 from 15 Swedish-speaking schools.

The general upper secondary education data consist of ME tests administered in spring and autumn 2015 and spring 2016. These are the tests in which the sample students from the 2013 national evaluation have most likely taken their matriculation test, provided that the regular duration for Finnish upper

secondary education is three years, less often four years, and that most candidates complete it in two rounds. In spring 2015, the ME test in English was taken by 21,336 candidates, in autumn 2015 by 3,405, and in spring 2016 by 25,853 candidates (Finnish Matriculation Examination Board, 2020).

Of the participants in the 2013 NELO tests, on average 51% opted for the general upper secondary education stream, while 41% preferred the vocational track (Härmälä, Huhtanen & Puukko, 2014). Of those students who participated in the NELO assessment, 1,485 could be tracked to the ME. The ME results were merged with the NELO data including the assessment results and the background questionnaires.

Table 1. A summary of participants' background (N = 1485)

		Frequency	Percent
Gender	Boy	645	44
	Girl	837	56
	Missing	3	0
Language of schooling	Finnish	1233	83
	Swedish	252	17
Parents' educational level	Both parents have taken the ME	418	28
	One of the parents has taken the ME	545	37
	Neither of the parents has taken the ME	387	26
	Missing	135	9

Measures

National Evaluation in English (NELO)

The basic education data were collected in spring 2013 as a paper-and-pencil test. The proficiency at the end of basic education was measured by tests of listening, reading and writing taken by all the sample pupils. The evaluation also comprised a speaking section. Only the tests of listening and reading were included in this study.

The composition of test tasks, themes, types and the intended CEFR level is depicted in Appendix 1. All the specifications were drawn from the National Core Curriculum for Basic Education (Finnish National Board of Education, 2004, henceforth NCC 2004) to ensure adequate pedagogical alignment. The target level for receptive skills (listening and reading) was B1.1. Before the main test, all the tasks and items were pilot tested by about one hundred pupils at schools of different kinds. These schools were not participating the main study of the national assessment. The pilot test covered a total of 71 items. Finally, 24 listening and 24 reading items were selected for the final main test alongside two anchor items from previous evaluations to establish comparability. One of the reading anchor items offered multiple-choice items in English, in contrast with the mainstream NELO policy to present questions and answer options in test-takers' L1. To define the difficulty of the proposed items, an item response theory (IRT) analysis was applied using the one-parameter Rasch model that locates the student latent ability and the item difficulty on the same continuum (de Ayala, 2009, pp. 4-16). Furthermore, teacher and pupil feedback on tasks with unusually high difficulty estimates or vaguely formulated instructions was considered in the final selection by editing and removing problematic items. After the main test, the IRT analysis was repeated (Härmälä, Huhtanen & Puukko, 2014.)

The solution percentages (i.e., percentages achieved of the maximum points) were calculated for each subskill for each student. The points achieved by the students were divided by the maximum score available and multiplied by a hundred. The maximum score for the listening test was 32 and for the reading test 28.

Matriculation Exam (ME)

The Matriculation Exam language test typically incorporates 30 listening and 30 reading items, a varied number of structure and vocabulary items and a writing task. In this study, only the listening and reading tests were used. The items for the ME are not pre-tested for confidentiality reasons. Following a tradition of openness, the tests are published on the Internet immediately after they have been administered. They are thus freely available for schools as training materials. Consequently, no item banking has been possible so far.

Instead, the quality of language tests of the ME Board is ensured by careful planning in alignment with the objectives in the national core curricula for each language syllabus. Each language specific section (English, Swedish, etc.) has a test construction group with expert sub-groups in charge of specific skills. They cross-check each other's item drafts in several rounds prior to the official reading in the language section that consists of representatives of all language subjects. Often, further revisions are proposed until the test is sent for a corrective check for its linguistic form, is translated to both national languages of instruction and ultimately processed for delivery.

The post-test analyses computed for ME items customarily include correlations across skill sections and in relation to the overall score. Based on these data, a few items may be removed from the final score calculation if bias is detected. The bias can be due to an ambiguous formulation of a question or related answer options. The final total score of 299 on language tests will then be shared on a norm basis into seven categories with a fail rate of about five per cent. The final distribution of grades varies between the tests being administered, because cut scores are based on the average of standardised total scores (Marjanen, 2015; Finnish Matriculation Examination website, 2020).

The maximum score for the listening test is 90, and that for the reading test 70. The test in English was digitalised in autumn 2018, but this study is based on the paper-and-pencil version. The solution percentages were computed for each subskill of the ME, as they were calculated for the NELO. Table 2 provides descriptive statistics for the proficiency in the NELO and ME for the subskills of listening and reading.

Table 2. Descriptive statistics for students' proficiency in the NELO and ME for the subskills of listening and reading

	M	SD	Skewness	Kurtosis
NELO, Listening	65.92	18.86	−.98	3.95
NELO, Reading	74.83	21.23	−1.27	4.39
ME, Listening	70.48	16.26	−.66	2.94
ME, Reading	72.02	16.80	−.56	2.53

All the analysed ME tests from spring 2015, autumn 2015 and spring 2016 share the same task structure. After listening to and reading a text in English, students completed 25 multiple-choice items in English and five constructed response items in the language of instruction, to avoid construct-irrelevant variance due to producing written text in the target language. The target level for advanced syllabus English is B2.1 in the CEFR (Finnish National Board of Education, 2003), which implies that most of the test items are targeted at this level, and a quarter of them below or above B2.1.

The aim of the test tasks is to reflect the themes and text types of the upper secondary courses, which, according to the operative curriculum of the test administrations in 2015 and 2016, incorporated the following course titles of compulsory (1-6) and optional courses (7-8):

(1) Young people and their world
(2) Communication and spare time
(3) Study and work
(4) Society and the world around us
(5) Culture
(6) Science, economy and technology
(7) Nature and sustainable development
(8) Our world and globalisation

(Finnish National Board of Education, 2003)

Data analysis

To explore RQ1 (i.e., the relationship between language proficiency and the background variables), we used regression analysis. A linear regression model (ordinary least-squares, OLS) was used to analyse the above-mentioned relationship at the mean of the score distribution. However, we were interested in the relationship not only on average but also at other points of the distribution. Thus, we analysed the relationship between the language proficiency and the background variables at the 0.25, 0.50 and 0.75 quantiles in the distribution by using quantile regression (QR) analysis (Koenker & Bassett, 1978). In QR, the relationship between the response variable and explanatory variables at different

quantiles of the conditional distribution are analysed instead of the average relationship which is examined in OLS. Thus, QR enables us to study whether explanatory variables predict language proficiency differently along the distribution — that is to say, among low-performing and high-performing students. In addition, the advantage of QR is that it is a more robust method and is less sensitive to outliers than OLS (Koenker & Bassett, 1978).

To explore RQ2, we conducted OLS and QR analysis to explore the extent to which the NELO predicts the language proficiency in the ME at the mean and at the 0.25, 0.50 and 0.75 quantiles in the distribution. We applied two models. In Model 1, only students' proficiency as indicated by their scores in the NELO predicted that on the ME. In Model 2, background variables (i.e., gender, parents' educational level, and language of instruction) were added to the analysis.

Results

To examine the relationship between the language proficiency and background variables (RQ1), we carried out regression analyses (OLS and QR). In the analyses, we investigated the effect of students' gender, language of instruction and parents' educational level on their performance in the NELO and ME. The results of OLS and QR for listening and reading are presented in Tables 3 and 4, respectively. The analysis results of listening and reading for the NELO were quite similar. The coefficients of background variables were larger at the 25th percentile than at the 75th percentile of the distribution, meaning that the effect of background variables was stronger for low-performing students than high-performing students.

Although girls' language proficiency as indicated by their test scores in the NELO seemed to be slightly lower than boys' language proficiency, the gender coefficient was statistically significant only at the median for listening, and at the median and 25th percentile for reading (see Tables 3 and 4). Students in Swedish language instruction schools performed better than students in Finnish language instruction schools. The coefficients were largest in the 25th percentile, whereas at the 75th percentile, language of instruction had no effect on the language proficiency. Moreover, the parents' educational level influenced the language

proficiency along all the distribution, suggesting that students with better educated parents also performed better. However, this effect was the most noticeable for low-performing students, indicating that for these students, the negative effect of parents' low educational level was more influential.

The effects of background variables on the language proficiency for the ME were almost the same as the NELO. Except for the effect of parents' educational level on the language proficiency in listening, the coefficients were more considerable for low-performing students than high-performing students. In listening, however, the coefficient of parents' educational level was the highest at the 75th percentile (i.e., for high-performing students), which differed from the NELO. In addition, different from what we found about the NELO, the gender coefficient was statistically significant along the distribution of the language proficiency in the case of the ME.

Table 3. The effect of background variables on listening proficiency (N = 1350)

	NELO				**ME**			
	OLS β (SE) [CI, 95 %]	**0.25** β (SE) [CI, 95 %]	**0.50** β (SE) [CI, 95 %]	**0.75** β (SE) [CI, 95 %]	**OLS** β (SE) [CI, 95 %]	**0.25** β (SE) [CI, 95 %]	**0.50** β (SE) [CI, 95 %]	**0.75** β (SE) [CI, 95 %]
Gender (Ref. boys)								
Girls	−1.31^{ns} (0.97) [−3.23, 0.60]	−3.13^{ns} (1.63) [−6.32, 0.07]	−3.13** (1.16) [−5.54, −0.86]	−1.56^{ns} (1.02) [−3.56, 0.44]	−3.93*** (0.87) [−5.64, −2.22]	−5.56*** (1.38) [−8.27, −2.84]	−2.78* (1.24) [−5.21, −0.34]	−2.78** (0.93) [−4.60, −0.95]
Language of instruction (Ref. Finnish-language)								
Swedish-language	6.36*** (1.26) [3.89, 8.82]	9.38*** (2.10) [5.25, 13.50]	6.25*** (1.49) [3.33, 9.17]	1.56^{ns} (1.31) [−1.02, 4.14]	5.47*** (1.12) [3.27, 7.68]	8.33*** (1.78) [4.84, 11.83]	5.56** (1.60 [2.42, 8.70]	2.78* (1.20) [0.43, 5.13]
Parents' educational level (Ref. Neither of the parents has taken the ME)								
One of the parents has taken ME	6.05*** (1.17) [3.76, 8.34]	6.25** (1.96) [2.41, 10.09]	9.38*** (1.39) [6.66, 12.09]	3.13* (1.22) [0.73, 5.52]	4.31*** (1.05) [2.25, 6.36]	2.78^{ns} (1.66) [−0.47, 6.03]	2.78^{ns} (1.49) [−0.14, 5.70]	5.56*** (1.11) [3.37, 7.74]
Both parents have taken ME	11.92*** (1.25) [9.48, 14.37]	15.63*** (2.08) [11.53, 19.72]	12.50*** (1.48) [9.60, 15.40]	7.81*** (1.30) [5.26, 10.37]	9.42*** (1.12) [7.24, 11.61]	8.33*** (1.77) [4.87, 11.80]	8.33*** (1.59) [5.22, 11.45]	11.11*** (1.19) [8.78, 13.44]
Constant	59.85*** (1.09) [57.70, 61.99]	50.0*** (1.83) [46.41, 53.59]	62.5*** (1.30) [59.96, 65.04]	76.56*** (1.14) [74.32, 78.81]	67.18*** (0.98) [65.26, 69.10]	58.33*** (1.55) [55.29, 61.38]	69.44*** (1.39) [66.71, 72.18]	77.78*** (1.04) [75.73, 79.82]
R^2 (OLS), Pseudo-R^2 (QR)	0.09	0.06	0.06	0.04	0.09	0.06	0.05	0.04

Note. The table presents estimated coefficients for linear regression (OLS) and quantile regressions from the 25th, 50th and 75th percentile of the distribution.

*p ≤ .050; **p ≤ .010; ***p ≤ .001; n.s. p > .050

Table 4. The effect of background variables on reading proficiency (N = 1350)

	NELO				**ME**			
	OLS β (SE) [CI, 95 %]	**0.25** β (SE) [CI, 95 %]	**0.50** β (SE) [CI, 95 %]	**0.75** β (SE) [CI, 95 %]	**OLS** β (SE) [CI, 95 %]	**0.25** β (SE) [CI, 95 %]	**0.50** β (SE) [CI, 95 %]	**0.75** β (SE) [CI, 95 %]
Gender (Ref. boys)								
Girls	−1.68^{ns} (1.10) [−3.84, 0.47]	−3.57* (1.78) [−7.06, −0.09]	−3.57** (1.18) [−5.89, −1.26]	0.00^{ns} (1.12) [−2.21, 2.21]	−3.55*** (0.89) [−5.30, −1.80]	−5.71** (1.67) [−8.99, −2.44]	−4.29** (1.27) [−6.77, −1.80]	−2.86** (0.90) [−4.62, −1.09]
Language of instruction (Ref. Finnish-language)								
Swedish-language	5.20*** (1.42) [2.42, 7.98]	7.14** (2.29) [2.65, 11.64]	3.57* (1.52) [0.59, 6.56]	0.00^{ns} (1.45) [−2.85, 2.85]	4.65*** (1.15) [2.39, 6.91]	5.71** (2.16) [1.49, 9.94]	4.29** (1.63) [1.08, 7.49]	2.86* (1.16) [0.58, 5.13]
Parents' educational level (Ref. Neither of the parents has taken the ME)								
One of the parents has taken ME	5.78*** (1.32) [3.20, 8.37]	10.71*** (2.13) [6.54, 14.89]	7.14*** (1.41) [4.37, 9.92]	3.57** (1.35) [0.93, 6.22]	3.22** (1.07) [1.12, 5.32]	5.71** (2.00) [1.78, 9.64]	4.29** (1.52) [1.31, 7.26]	2.86** (1.08) [0.74, 4.97]
Both parents have taken ME	12.45*** (1.40) [9.70, 15.21]	17.86*** (2.27) [13.40, 22.31]	10.71*** (1.51) [7.76, 13.67]	7.14*** (1.44) [4.32, 9.96]	10.35*** (1.14) [8.11, 12.59]	14.29*** (2.14) [10.09, 18.48]	12.86*** (1.62) [9.68, 16.03]	8.57*** (1.15) [6.31, 10.83]
Constant	69.27*** (1.23) [66.85, 71.69]	57.14*** (1.99) [53.23, 61.05]	75.0*** (1.32) [72.40, 77.60]	85.71*** (1.26) [83.24, 88.19]	68.81*** (1.00) [66.85, 70.78]	57.14*** (1.88) [53.46, 60.82]	70.0*** (1.42) [67.21, 72.79]	82.86*** (1.01) [80.88, 84.84]
R^2 (OLS), Pseudo-R^2 (QR)	0.07	0.07	0.04	0.02	0.09	0.06	0.06	0.04

Note. The table presents estimated coefficients for linear regression (OLS) and quantile regressions from the 25th, 50th and 75th percentile of the distribution.

*p ≤ .050; **p ≤ .010; ***p ≤ .001; n.s. p > .050

In addition to the effect of background variables, we were interested in the relationship between the language proficiency as indicated by test scores on the NELO and that on the ME (RQ2). The OLS and QR were carried out in two phases. First, only the NELO was used as a predictor to the model. Second, background variables were added in the model. The results for listening and reading are presented in Tables 6 and 7, respectively. The results indicate that the test scores on the NELO predict these on the ME differently at different points of the distribution. For both subskills, the coefficients were positive and statistically significant at both the lower and upper ends of the distribution. However, the coefficients were the largest at the lower part of the distribution, and students' scores on the NELO explained more variance of their scores on the ME for low-achieving students than for high-achieving students. The effect of students' scores on the NELO on their performance on the ME was almost the same after the background variables were added to the model. In other words, students' scores on the NELO and background variables together explained only slightly more variance than their scores on the NELO alone (see Tables 5 and 6).

Table 5. The effect of students' NELO listening scores and background variables on their ME listening scores

	OLS β (SE) [CI, 95 %]	**0.25** β (SE) [CI, 95 %]	**0.50** β (SE) [CI, 95 %]	**0.75** β (SE) [CI, 95 %]
NELO	.47*** (.02) [.44, .51]	.74*** (.02) [.69, .79]	.61*** (.02) [.57, .66]	.42*** (.02) [.38, .47]
Constant	39.32*** (1.28) [36.80, 41.84]	12.96*** (1.65) [9.73, 16.20]	30.90*** (1.56) [27.85, 33.96]	51.75*** (1.60) [48.62, 54.89]
R^2 (OLS), Pseudo-R^2 (QR)	0.30	0.27	0.20	0.13
N	1,485	1,485	1,485	1,485
NELO	.48*** (.02) [.44, .52]	.71*** (.03) [.66, .77]	.63*** (.03) [.58, .68]	.44*** (.03) [.40, .49]
Gender (Ref. boys)				
Girls	−3.30*** (.74)	-2.78** (1.02)	-2.67** (.90)	−1.39^{ns} (.90)

	[−4.74, −1.85]	[−4.78, −.77]	[−4.44, −.91]	[−3.15, .37]
Language of instruction (Ref. Finnish-language)				
Swedish-language	2.42* (.96) [.54, 4.31]	2.22^{ns} (1.33) [−.39, 4.83]	1.95^{ns} (1.17) [−.35, 4.26]	1.39^{ns} (1.17) [−.90, 3.68]
Parents' educational level (Ref. Neither of the parents has taken the ME)				
One of the parents has taken ME	1.41^{ns} (.89) [−.34, 3.16]	1.67^{ns} (1.24) [−.76, 4.09]	2.06^{ns} (1.09) [−.08, 4.20]	1.39^{ns} (1.09) [−.74, 3.52]
Both parents have taken ME	3.71*** (.97) [1.80, 5.62]	3.33* (1.35) [.68, 5.98]	4.01** (1.19) [1.68, 6.35]	2.78* (1.19) [.45, 5.11]
Constant	38.48*** (1.49) [35.57, 41.40]	14.44*** (2.06) [10.40, 18.49]	28.50*** (1.82) [24.93, 32.06]	48.61*** (1.81) [45.06, 52.16]
R^2 (OLS), Pseudo-R^2 (QR)	0.35	0.29	0.24	0.16
N	1,350	1,350	1,350	1,350

Note. The table presents estimated coefficients for linear regression (OLS) and quantile regressions from the 25th, 50th and 75th percentile of the distribution.
**p ≤ .050; **p ≤ .010; ***p ≤ .001; n.s. p > .050*

Table 6. The effect of students' NELO reading scores and background variables on their ME reading scores

	OLS β (SE) [CI, 95 %]	**0.25** β (SE) [CI, 95 %]	**0.50** β (SE) [CI, 95 %]	**0.75** β (SE) [CI, 95 %]
NELO	0.41*** (0.02) [0.38, 0.45]	0.66*** (0.02) [0.62, 0.70]	0.58*** (0.02) [0.54, 0.62]	0.38*** (0.02) [0.33, 0.42]
Constant	41.13*** (1.36) [38.46, 43.81]	13.61*** (1.70) [10.27, 16.96]	28.83*** (1.69) [25.52, 32.14]	53.38*** (1.81) [49.83, 56.94]
R^2 (OLS), Pseudo-R^2 (QR)	0.27	0.25	0.20	0.11
N	1,485	1,485	1,485	1,485
NELO	0.43*** (0.02) [0.39, 0.46]	0.63*** (0.03) [0.58, 0.68]	0.58*** (0.02) [0.53, 0.63]	0.38*** (0.03) [0.32, 0.43]
Gender (Ref. boys)				

Girls	−2.83*** (0.76) [−4.32, −1.34]	−1.43^{ns} (1.03) [−3.45, 0.60]	−2.60** (0.97) [−4.50, −0.69]	−2.86* (1.11) [−5.03, −0.68]
Language of instruction (Ref. Finnish-language)				
Swedish-language	2.43* (0.99) [0.50, 4.37]	2.45^{ns} (1.34) [−0.17, 5.07]	2.86** (1.26) [0.39, 5.32]	2.71^{ns} (1.44) [−0.11, 5.52]
Parents' educational level (Ref. Neither of the parents has taken the ME)				
One of the parents has taken ME	0.76^{ns} (0.92) [−1.05, 2.56]	0.61^{ns} (1.25) [−1.83, 3.06]	0.52^{ns} (1.17) [−1.78, 2.82]	0.60^{ns} (1.34) [−2.02, 3.23]
Both parents have taken ME	5.04*** (1.00) [3.07, 7.00]	5.92*** (1.36) [3.26, 8.58]	3.90** (1.28) [1.39, 6.40]	3.61* (1.46) [0.75, 6.47]
Constant	39.27*** (1.56) [36.21, 42.33]	14.69*** (2.12) [10.54, 18.85]	28.31*** (1.99) [24.41, 32.22]	52.78*** (2.28) [48.32, 57.25]
R^2 (OLS), Pseudo-R^2 (QR)	0.34	0.29	0.24	0.15
N	1,350	1,350	1,350	1,350

Note. The table presents estimated coefficients for linear regression (OLS) and quantile regressions from the 25th, 50th and 75th percentile of the distribution.

**p ≤ .050; **p ≤ .010; ***p ≤ .001; n.s. p > .050*

While students' test scores on the ME were predicted by both their test scores on the NELO and background variables, the gender coefficient was statistically significant in the 25th percentile and in the median for listening, whereas for reading, it was statistically significant at the median and at the 75th percentile but not at the 25th percentile. The results of the effect of background variables mentioned above showed that the effect of the language of instruction on the test scores on the ME (in favour of Swedish language instruction schools) was statistically significant along the distribution. However, while the test scores on the ME were explained by both the test scores on the NELO and background variables, the effect of language of instruction was not statistically significant at the 25th or 75th percentile.

Discussion

Among the multiple facets of the meaningfulness framework proposed by Kunnan (2018), our study addressed the principle of fairness through the

predictive power of a test compared to administration of a subsequent test with the same group of test-takers. RQ1 focuses on the capacity of compulsory basic education to produce equal learning outcomes with respect to gender, parental educational level and language of schooling. The findings reveal the uneven language proficiency attained by boys and girls, which persists across general upper secondary education. Similar differences prevail in favour of students with better educated parents and from schools in which the language of instruction is Swedish. These results affirm the predictive validity of the NELO, thereby warranting the overall reliability of the assessment and thereby strengthening the fairness of the test itself at a general level.

Research on the predictive validity of the language tests has mostly been focused on the average achievement of the students. The strength of this study is that it examines the predictive validity along the distribution. At a closer glance, the findings suggest that the association between the proficiency at the end of basic education and at the end of upper secondary education is stronger for low-achieving students than those with higher proficiency. These findings can be considered as rebuttals of fairness from the test consistency aspect of Kunnan's framework, because both groups have completed similar upper secondary education based on the national core curriculum, but in fact, we are dealing with a problem of social justice. Inequal learning outcomes between background variables, such as gender, parents´ education and language of school instruction deserve ongoing attention in educational policymaking.

In contrast to many other evaluations of literacy skills suggesting the superiority of girls, the results of our study, in the case of both the NELO and ME subtests of receptive skills clearly attested to boys' higher proficiency compared to girls. The finding is consistent with the findings from national assessments of English in Finland (Tuokko, 2007) and Sweden (Börjesson & Nilsson, 2018). Since most of the items in receptive tests adopt the multiple-choice format, this might suggest that this item format favours male students, and that they are more inclined to risk-taking in test situations than their female counterparts (Karimi & Biria, 2017). On the other hand, some earlier studies (e.g., Olsen et al., 2001) highlighted that a single item characteristic such as the response type has no systematic effects, but the students' responses are affected by complex

interactions between item characteristics (e.g., item wording and item format).

As stated above, the advantage of boys over girls is a particularity of English language assessments, probably due to the large proportion of proficiency that has been acquired outside school by gaming and the use of digital media, which necessitates language use. This phenomenon has been noted in other countries as well (Sundqvist & Sylvén, 2016; Finnish Ministry of Education and Culture, 2022; OECD, 2019). Consequently, the earlier results of the NELO showed an insignificant association between teaching practices and student performance for the entire cohort (Härmälä, Huhtanen & Puukko, 2014). However, more research is needed on the factors through which the effect of gender and other background variables on language proficiency are mediated. Although the sub-principles of fairness of Kunnan's model (presented above in the validity section) have formally been met, by general right and provision of school education and adequate opportunity of taking the test, there are more subtle sources of inequality deriving from sub-cultures that may pose diverse expectations on educational success for boys and girls. It is also test developers' responsibility to ensure a balanced and diverse selection of task types that do not have a gendered bias.

Although boys outperformed girls in both assessments, the effect of gender on language proficiency seemed to be greater in the case of ME than NELO. On the one hand, this might indicate that upper secondary education cannot level out the difference between genders. On the other hand, it is worth noting that the assessments were not comparable regarding their stakes for the students, which may have affected their results (Knekta, 2017). In the case of ME, the test-takers would certainly do their best as they were fully aware of the high stakes attached to in the exam. As for NELO, teachers are mandated to administer the national evaluation, but its results may or may not be considered in grading the pupils. The stakes are not high, which the pupils know.

The finding indicating the superiority of students with parents with a higher educational level is hardly surprising since the effect of socio-economic status on students' academic achievement is a well-known fact (e.g., Sirin 2005). However, this study suggests that this effect is more considerable for low-achieving

students than their higher-achieving counterparts, especially at the end of basic education. This also applies to the results related to the effect of language of instruction since students in Swedish language instruction schools were found to outperform those in Finnish language instruction schools. A plausible explanation is that Swedish and English languages are linguistic cognates, while the Finnish language belongs to a different language family. Nevertheless, these findings pose a challenge to the ethos of equality and equity attributed to Finnish education. In particular, it seems that greater efforts are needed from policy-makers to provide more educational support to low-performing pupils from less advantaged backgrounds. It is imperative that fairness and justice prior to the test be catered for, because inequalities exist in the society beyond the test itself. At the level of test design, item writers should ensure a rich variation of topic coverage to match the realities of multiple test-taker populations.

Conclusion

The first research question addressed the relationship between the receptive language proficiency and background variables (i.e., gender, language of instruction and parents' educational level). The effects of background variables on language proficiency were almost the same for the two tests in focus. Boys performed slightly better than girls on the NELO, but when it comes to the ME, the gender effect in favour of boys was more prominent across the skills and level distribution. Students in Swedish language instruction schools performed better than those in Finnish language schools for both tests. Parents' educational level influenced the language proficiency along all the distribution, suggesting that students with better educated parents performed better. Overall, the effect of the background variables was stronger for low-performing students than for high-performing students. For the second research question, students' scores on the NELO and their background variables were used in the model to predict their test performance on the ME. For low-achieving students, their scores on the NELO explained more variance in their performance on the ME, as compared with their high-achieving peers. This trend was the applicable to both listening and reading. The effect of students' scores on the NELO on their performance on the ME was almost the same, if the background variables were included in the model.

We acknowledge a few limitations with this study. First, the target group consisted only of general upper secondary students who took the ME. There is no general school-leaving exam for students in the vocational upper secondary tracks, and these students comprised 41% of the 2013 cohort. We might also assume that the students in the general track are a more selected group because of the higher grade-point average required for admission into general upper secondary schools. Therefore, the knowledge gained on the predictive power of the NELO is incomplete and applies only to the students in the general upper secondary track. It is also worth reiterating that in the current study we could not control for any factors related to upper secondary instruction at the local level, such as the quality of instruction, group size or provision and completion of courses at individual schools.

Furthermore, a more fine-grained account of the performance of the multiple groups might have been achieved by comparing their test results on several item types. Multiple-choice and constructed response questions entail different sub-skills and divergent strategies deployed differently by test takers of different genders (In'nami & Koizumi, 2009; Karimi & Biria, 2017). A more in-depth analysis into the content of the listening and reading tasks could be performed by applying differential item functioning (DIF) analysis that allows for investigation of single test items with respect to their differential power (measuring different abilities for members of separate subgroups) and other relevant features. This line of research has been initiated for the Finnish ME items (von Zansen et al., 2022). To incorporate more advanced item analyses as regular steps in test validation, policy-makers should equip test developers with sufficient time and other resources to enhance the fairness and justice of nation-wide assessments.

Despite the reasonably attested meaningfulness of the NELO as a predictor of future achievement in receptive language abilities, certain concerns can be raised about the differentiated learning outcomes determined by gender, language of schooling and parental educational level. There are gaps to be bridged at both the basic compulsory level and in upper secondary education to meet the recognised challenges by encouraging lower-performing students, especially girls and students of lower socio-economic status, in pursuit of equal gain of studying languages to enable full participation in society.

This study merely addressed receptive skills that were more comparable in the methodological sense than the sections for productive skills, in which the rating scales differed substantially in the two assessments. For future scrutiny, serious attention should be paid to the symmetry of illustrative proficiency-based rating scales and reporting routines to enable reliable and consistent longitudinal accounts of larger student populations' communicative language skills across educational trajectories.

To promote the overall fairness of national assessments and to detect longitudinal trends, there is a need for well-designed replication studies across several administrations of both assessments. There are limited opportunities to do this, because the NELO is carried out less frequently than the ME, which is held twice a year. However, the interval between the NELO administrations is diminishing, enabling more frequent and timely comparisons (Härmälä et al., 2019). A tangible improvement to follow worldwide trends in foreign and second language proficiency is the envisaged introduction of a foreign language test into the PISA programme in 2025 (OECD, 2022).

References

de Ayala, R.J. (2009). *The theory and practice of item response theory*. The Guilford Press.

Bachman, L.F., & Palmer, A.S. (2010). *Language Assessment in Practice.* Oxford University Press.

Beuchert, L., & Nandrup, A. (2017). *The Danish national tests at a glance.* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2996431

Börjesson, L., & Nilsson, S. (2018). Könsskillnader i engelsk språkfärdighet. In G. Erickson (Ed.), *Att bedöma språklig kompetens* (pp. 107–118). Rapporter från projektet Nationella prov i främmande språk. RIPS: Rapporter från Institutionen för pedagogik och specialpedagogik, nr 16. Göteborgs universitet. https://gupea.ub.gu.se/bitstream/handle/2077/57683/gupea_2077_57683_2.pdf?sequence=2

Council of Europe (2001). *Common European framework of reference for languages: learning, teaching, assessment*. Cambridge University Press. https://www.coe.int/en/web/common-european-framework-reference-languages

European Commission (2012). *First European survey on language competences. Final report*. https://education.ec.europa.eu/ Retrieved 10.2.2022.

Equality Act. (2014). 5 § (30.12.2014/1329) Finland's Ministry of Justice. https://finlex.fi/fi/laki/ajantasa/1986/19860609#a30.12.2014-1329

Equity Act. (2014). 6§, 8§. 1325/2014. Finland's Ministry of Justice. https://www.finlex.fi/fi/laki/alkup/2014/20141325

Finnish Ministry of Education and Culture. (2022, November 20). *Finnish Matriculation Examination Board website*. https://www.ylioppilastutkinto.fi/en/

Finnish National Board of Education. (2003). *National core curriculum for general upper secondary education 2003*. Finnish National Board of Education.

Finnish National Board of Education. (2004). *National core curriculum for basic education 2004*. Finnish National Board of Education.

Finnish National Board of Education. (2015). *National core curriculum for upper secondary schools 2015*. Finnish National Board of Education.

Finnish National Agency of Education (2022, July 7). *Equity and equity*. https://www.oph.fi/fi/tasa-arvo-ja-yhdenvertaisuus

Fischbach, A., Keller, U., Preckel, F., & Brunner, M. (2013). PISA proficiency scores predict educational outcomes. *Learning and Individual Differences*, 24, 63–72. https://www.sciencedirect.com/science/article/pii/S104160801200163X

Fulcher, G., & Davidson, F. (2007). *Language Testing and Assessment*. Routledge.

Hildén, R., & Rautopuro, J. (2014a). *Ruotsin kielen A-oppimäärän oppimistulokset perusopetuksen päättövaiheessa 2013*. [National evaluation of learning outcomes in Swedish language at the end of

compulsory basic education 2013]. Finnish Education Evaluation Centre. 2014:1. https://karvi.fi/publication/ruotsin-kielen-oppimaaran-oppimistulokset-perusopetuksen-paattovaiheessa-2013/

Hildén, R., & Rautopuro, J. (2014b). *Saksan kielen A- ja B-oppimäärän oppimistulokset perusopetuksen päättövaiheessa 2013*. [National evaluation of learning outcomes in German language at the end of compulsory basic education 2013. Finnish Education Evaluation Centre 2014:4. https://karvi.fi/publication/saksan-kielen-ja-b-oppimaaran-oppimistulokset-perusopetuksen-paattovaiheessa-2013/

Hildén, R., & Takala, S. (2007). Relating Descriptors of the Finnish School Scale to the CEF Overall Scales for Communicative Activities. In Koskensalo, A., J. Smeds, P.,Kaikkonen, & V. Kohonen (Eds.) *Foreign languages and multicultural perspectives in the European context; Fremdsprachen und multikulturelle Perspektiven im europäischen Kontext*. Dichtung, Wahrheit und Sprache. LIT-Verlag, 291 - 300.

Härmälä, M., & Huhtanen, M. (2014). Ranskan *kielen A- ja B-oppimäärän oppimistulokset perusopetuksen päättövaiheessa 2013***.** [National evaluation of learning outcomes in French language at the end of compulsory basic education 2013. Finnish Education Evaluation Centre 2014:3. https://karvi.fi/publication/ranskan-kielen-ja-b-oppimaaran-oppimistulokset-perusopetuksen-paattovaiheessa-2013/

Härmälä, M., Huhtanen, M., & Puukko, M. (2014). *Englannin kielen A-oppimäärän oppimistulokset perusopetuksen päättövaiheessa 2013* [Learning outcomes in English advanced syllabus at the end of basic education]. Finnish National Board of Education and Finnish Education Evaluation Centre 2014:2. https://karvi.fi/publication/englannin-kielen-oppimaaran-oppimistulokset-perusopetuksen-paattovaiheessa-2013/

Härmälä, M., Huhtanen, M., Puukko, M., & Marjanen, J. (2019). *A-englannin oppimistulokset 7. luokan alussa* [Learning outcomes in advanced syllabus English in the beginning of grade 7 of basic education]. National Evaluation Centre. Publications 13:2019. https://karvi.fi/publication/a-englannin-oppimistulokset-7-luokan-alussa-2018/

Härmälä, M., Huhtanen, M., Silverström, C., Hilden, R., Rautopuro, J., & Puukko, M. (2014). Inlärningsresultaten I främmande språk i de svenskspråkiga skolorna 2013. A-lärokursen i engelska samt B-lärokurserna i franska, tyska och ryska. [Learning outcomes in foreign languages in Swedish instruction schools 2013. Advanced syllabus in English and short syllabi in French, German and Russian]. National Board of Education, Finnish Education Evaluation Centre 2014:6. https://www.oph.fi/sites/default/files/documents/160083_inlarningsresultaten_i_frammande_sprak_i_de_svensksprakiga_skolorna_20131.pdf

In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing*, *26*(2), 219-244.

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17–64). American Council of Education and Praeger.

Kane, M. (2012). Validating score interpretations and uses. *Language Testing*, *29*(1), 3–17.

Karimi, M. & Biria, R. (2017). Impact of risk taking strategies on male and female EFL learners' test performance: The case of multiple-choice questions. *Theory and Practice in Language Studies*, *7*(10), 892. https://www.academypublication.com/issues2/tpls/vol07/10/10.pdf

Knekta, E. (2017). Are all pupils equally motivated to do their best on all tests? Differences in reported test-taking motivation within and between tests with different stakes. *Scandinavian Journal of Educational Research*, *61*(1), 95-111. https://doi.org/10.1080/00313831.2015.1119723

Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. *Econometrica*, *46*(1), 33–50

Korp, H. (2006). *Lika chanser i gymnasiet? En studie om betyg, nationella prov och social reproduktion*. Lärarutbildningen, Malmö högskola.

Kunnan, A. J. (2018). *Evaluating language assessments*. Taylor & Francis.

Leino, K., & Nissinen, K. (2018). Suomalaisoppilaiden lukemiseen sitoutuminen, taustatekijät ja lukutaito: yhteyksien etsiminen

polkuanalyysillä. [Finnish pupils´ engagement in reading, background factors and reading skills: finding connections with path analysis]. In J. Rautopuro, & K. Juuti (Eds.), PISA pintaa syvemmältä: PISA 2015 Suomen pääraportti [PISA deeper than the surface. PISA 2015 Main report of Finland] (pp. 39-67). Suomen kasvatustieteellinen seura. Kasvatusalan tutkimuksia, 77. http://urn.fi/URN:ISBN:978-952-5401-82-0

Leino, K., Rautopuro, J. & Kulju, P. (2021). Esipuhe. [Foreword]. *Lukutaito – Tie tulevaisuuteen: PISA 2018 Suomen pääraportti*. [Literacy – Road to the future: PISA 2018. Main report of Finland] (pp. 7- 12). The Finnish Educational Research Association http://urn.fi/URN:ISBN:978-952-7411-16-2

Marjanen, J. (2015). *Ylioppilasarvosanojen vertailukelpoisuusongelma ja SYK-menetelmän edellytykset sen ratkaisemiseen*. [The problem of comparability of Matriculation Examination scores and the conditions of the average of standardized scores to solve it] *Kasvatus, 46 (6), 317-333*.

Matějů, P., & Smith, M. L. (2015) Are boys that bad? Gender gaps in measured skills, grades and aspirations in Czech elementary schools. *British Journal of Sociology of Education 36*(6), 871–895. https://www.tandfonline.com/doi/full/10.1080/01425692.2013.874278

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education and Macmillan.

Murdoch, J., Kamanzi, P. C., & Doray, P. (2011). The influence of PISA scores, schooling and social factors on pathways to and within higher education in Canada. *Irish Educational Studies*, *30*(2), 215–235. https://www.tandfonline.com/doi/full/10.1080/03323315.2011.569142

OECD. (2019). *PISA 2018 Results* (Volume II): *Where all students can succeed - Girls' and boys' performance in PISA*. https://doi.org/10.1787/f56f8c26-en

OECD (2022, February 20). *Pisa website*. https://www.oecd.org/pisa/foreign-language/

Olsen, R.V., Turmo, A. & Lie, S. (2001). Learning about students' knowledge and thinking in science through large-scale quantitative studies. *European Journal of Psychology of Education 16*, 403 - 420. https://doi.org/10.1007/BF03173190

Ouakrim-Soivio, N., Rautopuro, J., & Hildén, R. (2018). Shadows under the north star- The inequality developing in Finnish school education. *Nordidactica* 2018 (3), 65 – 86. https://journals.lub.lu.se/nordidactica/article/view/19073/17262

Pizorn, K., & Huhta, A. (2016). Assessment in educational settings. In Tsagari, D. & Banerjee, J. (Eds.) *Handbook of Second Language Assessment* (pp. 239–254). Mouton De Gruyter.

Pulkkinen, J., & Rautopuro, J. (2022). The correspondence between PISA performance and school achievement in Finland. *International Journal of Educational Research,* 114, Article 102000. https://doi.org/10.1016/j.ijer.2022.102000

Rosdahl, A. (2014). *Fra 15 år til 27 år: PISA 2000-eleverne i 2011*/12 [The PISA 2000 students in 2011/12]. SFI-Det Nationale Forskningscenter for Velfærd.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of educational research*, *75*(3), 417-453.

Statistics Finland. (2020). *Appendix table 1. Students and completed qualifications and degrees in education leading to a qualification or degree in 2020*. https://docs.google.com/spreadsheets/d/1UmH1h827y4JMlzzGcAklgn6rVOnIrMnkpb6Uz8wgTU0/edit#gid=1954000653

Sundqvist, P., & Sylvén, L. K. (2016). *Extramural English in Teaching and Learning: From Theory and Research to Practice*. Springer.

Toropainen, O. (2010). Utvärdering av läroämnet finska i den grundläggande utbildningen. *Inlärningsresultaten i finska enligt A-lärokursen och den modersmålsinriktade lärokursen i årskurs*, 9. [Learning outcomes in English advanced syllabus at the end of basic education]. Publications 2010:1. Finnish National Board of education.

Toulmin, S. E. (1958). *The Uses of Argument*. Cambridge University Press.

Tuokko, E. (2007). *Mille tasolle perusopetuksen englannin opiskelussa päästään? Perusopetuksen päättövaiheen kansallisen arvioinnin 1999 Eurooppalaisen viitekehyksen taitotasoihin linkitetyt tulokset*. Jyväskylän yliopisto. [What level do pupils reach in English at the end of the comprehensive education? National assessment results in 1999 linked to the Common European Framework.] Jyväskylä Studies in Humanities. University of Jyväskylä Dissertations Archive. https://jyx.jyu.fi/bitstream/handle/123456789/13426/9789513927677.pdf?sequence=1&isAllowed=y

von Zansen, A., Hilden, R., & Laihanen, E. (2022). The multimodal listening test in a high-stakes context: Gender-neutral or not? *International Journal of Listening, 1-18*. https://www.researchgate.net/publication/358075666_The_Multimodal_Listening_Test_in_a_High-Stakes_Context_Gender-Neutral_or_not

Weideman, A. (2019). Degrees of adequacy: tTe disclosure of levels of validity in language assessment. *Koers*, 84(1), 1–15. https://doi.org/10.19108/KOERS.84.1.2451

Appendix 1. Test specification NELO 2013 (Härmälä, Huhtanen & Puukko, 2014)

	Title	Theme in curriculum	CEFR level	N and type of items
L1	Rules of a game	spare time, hobbies	A1-A2	3 mc
L2	Report about Ayrton Senna	spare time, media	B1.1	3 mc
L3	Bonfire night celebration	culture	B2.1	3 mc
L4	Dialogues	public service pets travel	A2	3 mc
L5	Announcements on the plane	travel meals	A1-A2	3 mc
L6	Weather forecast	nature	A2-B1	3 cr
L7	Tips for the weekend	travel culture	A1 A2-B1	3 cr
L7	Animal rescues	health and welfare spare time	B1	3 cr
R1	Wales	tourism culture	A2.1	3 mc
R2	Dressing advice for boys	everyday life shopping sustainable development	B1.1	3 mc
R3	Interview with Simon Cowell	working life media: music	B1.2	3 mc
R4	Short texts about recycling	sustainable development daily life	A2.2 B1.2	3 mcE
R5	A news item about a kangaroo	working life living in the country and in the city	A2-B1	3 cr
R6	A letter from the Mordock-Bowers family	family life sports	B1.1	3 cr
R7	A news item about a traffic accident	health and welfare daily life	A2-B1	3 cr
R8	Advice for a job interview	working life	B1.1	3 cr
W1	Message to a hotel reception (40 – 60 words)	public service spare time	B1	
W2	My favourite book/ film	culture: literature, films	A2 – B2	

Notes. L = Listening task; R = Reading task; W = Writing task; mc = multiple-choice item in the language of instruction; mcE= multiple-choice item in English; cr = constructed response in the language of instruction.