Journal of Second Language Writing 48 (2020) 100700

Contents lists available at ScienceDirect

Journal of Second Language Writing

journal homepage: www.elsevier.com/locate/jslw

English writing skills of students in upper secondary education: Results from an empirical study in Switzerland and Germany

Stefan D. Keller[a,*], Johanna Fleckenstein[a,b], Maleika Krüger[a], Olaf Köller[b], André A. Rupp[c]

[a] School of Teacher Education, University of Applied Sciences North-Western Switzerland, Hofackerstr. 30, 4132 Muttenz, Switzerland
[b] Leibniz Institute for Science and Mathematics Education, Olshausenstrasse 62, 24118 Kiel, Germany
[c] Educational Testing Service (ETS), 660 Rosedale Rd, Thurstone Building, Mailstop T-03, Princeton, NJ, 08540, USA

ARTICLE INFO

Keywords:
English as a foreign language
Essay writing
Upper secondary education
Longitudinal study

1. Introduction

Foreign language skills are vital for learners' active participation in modern knowledge societies and integration in an international democratic system (Keller, 2013; Rychen & Salganik, 2003). The Council of Europe (CoE) recommended that every citizen of the European Union should be able to use three languages at the end of secondary school in order to equip them for the challenges of intensified international mobility and closer co-operation not only in education, culture and science but also in trade and industry (CoE, 2008).

The Common European Framework of Reference for Languages (CEFR) was created to provide transparent level descriptors for language competences, which form the basis of educational standards in the European Union. These descriptors were originally scaled on a sample of about 2800 learners from Swiss lower and upper secondary, vocational, and adult education classrooms (CoE, 2001). Neither before nor after this work, however, has there a been representative study of learners' writing competences in Switzerland or Germany beyond Year 9 (i.e., after the end of compulsory schooling). There is thus a significant research gap concerning English as a foreign language (EFL) writing competences at upper secondary level in these two countries.

In the absence of reliable data, universities in Switzerland raised concerns regarding students' entry level English competences,

* Corresponding author.
E-mail addresses: stefan.keller@fhnw.ch (S.D. Keller), fleckenstein@ipn.uni-kiel.de (J. Fleckenstein), maleika.krueger@fhnw.ch (M. Krüger), koeller@ipn.uni-kiel.de (O. Köller), arupp@ets.org (A.A. Rupp).

https://doi.org/10.1016/j.jslw.2019.100700
Received 9 January 2019; Received in revised form 6 November 2019; Accepted 19 November 2019
Available online 09 January 2020
1060-3743/ © 2020 Published by Elsevier Inc.

particularly in writing (Brupbacher, Jucker, König, Roth, & Straumann, 2008). A report commissioned by a major Swiss university recommended that students should be trained in argumentative writing, in particular structuring paragraphs, formulating a thesis, weighing arguments and using appropriate rhetorical devices (Brupbacher et al., 2008).

The current study was designed to investigate these research gaps by focusing on the academic track of upper secondary education (i.e., general education grammar schools / *Gymnasium*). As this is the most selective type of school both in Germany and Switzerland, graduates typically choose demanding courses of study in tertiary education that require strong productive and receptive skills in English. The study focused on students in their penultimate year before graduation, where *argumentative writing* and *source-based writing* are two key genres which figure prominently in the relevant curricula (Fleckenstein, Keller, Krüger, Tannenbaum, & Köller, 2019).

By testing English writing in those two genres, the study investigated whether students achieved the overall target of upper secondary EFL writing (i.e., CEFR level B2). The study further sought to describe how students' writing competences develop over the course of approximately one school year. Lastly, it evaluated the relationship between individual factors (i.e., gender, language background, general cognitive ability, and socioeconomic status) and systemic factors (i.e., differences in the school systems across countries) that might drive different educational trajectories.

We have divided this paper into four main sections as follows. In the first section, we summarize the relevant background, focusing on the educational systems in the two countries and learning gains typically found in EFL education at upper secondary level. As the source-based writing prompts used in the study require students to understand and integrate written and auditory input material, we also report key findings from studies on receptive competences, i.e., reading and listening (Hartig & Jude, 2008; van Ockenburg, van Weijen, & Rijlaarsdam, 2016; Schoonen, 2019).

In the second section, we detail the methods of data-gathering and measures of writing employed as well as presenting an overview of scaling techniques and reliability indicators. This includes a short discussion of *automated essay evaluation* (AEE), which was an important feature of analysis in this study. As these aspects have been described in detail in a previous publication (Rupp, Casabianca, Krüger, Keller, & Köller, 2019), we only provide a summary here. In the third section, we describe key results for our research questions using descriptive statistics and regression models with moderating variables. In the final section, we discuss the results in terms of their implications for the future of writing education in participating countries and a wider context.

2. Background

2.1. Upper secondary school systems in Germany and Switzerland

In Germany and Switzerland, only a selective group of students goes on to upper secondary education (International Standard Classification of Education [ISCED] level 3), while the majority leave school after Year 9 or 10 (ISCED level 2) to do apprenticeships. Specifically, about 21% of students in Switzerland and about 35% in Germany completed upper secondary education in academic track schools in 2015 (National Statistical Office of Switzerland [BFS], 2016; Standing Conference of the Ministers of Education in Germany [KMK], 2018).

While the Swiss system is thus more restrictive than the German one, EFL curricula in both countries mandate similar competences to be taught in each school year, with argumentative and source-based writing skills particularly relevant for the last two years before graduation. In order for students to acquire the prerequisite linguistic skills and composition strategies, these complex writing skills must be given broad scope in the two final years for students to achieve the ambitious educational goals at graduation (Harsch, Schröder, & Neumann, 2008).

2.2. Educational standards in EFL

Educational standards in Germany and Switzerland, in line with most other European countries, set CEFR level B2 as the target educational goal for EFL competences at upper secondary level (European Commission, 2017). In Germany, B2 is the level of English which typical students are expected to reach while high performing students are expected to reach level C1 (Standing Conference of the Ministers of Education and Cultural Affairs in the Federal Republic of Germany KMK, 2014). The curriculum of Schleswig-Holstein, the German state included in this study, specifies that students should learn to write texts about a wide range of topics of their academic and personal interest in a way that is appropriate to topic and addressee (Institute for Quality Development in Schools of Schleswig Holstein [IQSH], 2014).

In Switzerland, which currently has no unifying national standards for foreign languages regulating its upper secondary schools, level B2 has become an unofficial target level for upper secondary education, with high performing learners also expected to reach level C1. For example, level B2 designates a "passing" to "good" grade in the canton of Berne (Educational Department of Berne [EDB], 2017). Similarly, the English curriculum in the canton of Basel states that students at the end of upper secondary education should be able to write clearly structured, argumentatively convincing longer texts (e.g., five-paragraph essays) (Educational Department of Basel-Stadt [EDBS], 2017). This language echoes descriptors at levels B2 and C1 of the CEFR overall writing production scale (CoE, 2001).

2.3. Studies of EFL writing competences in secondary education

Research on English writing at the secondary level in the European and international context is abundant (see, for example,

Rijlaarsdam et al., 2013; Schoonen, 2019; Schoonen, van Gelderen, Stoel, Hulsijn, & Glopper, 2010). Consequently, in this section we are concerned only with research relating to Germany and Switzerland where empirical studies are much scarcer. The largest data set concerning learners' EFL writing competences in Germany is from the study *German English Student Competences International* (DESI) (Harsch et al., 2008; Hartig & Jude, 2008; Klieme et al., 2008).

The DESI study assessed learners' ability to write reports and letters of advice in English at the end of Year 9 in a sample of about $n = 11{,}000$ students and showed that 62% of learners were able to write texts at or above level A2 of the CEFR, while only 4% reached level B2, which would allow them to write texts that are clear and detailed about various topics. Regarding different genres, DESI showed that learners found it easier to write in narrative genres (e.g., describing a person) than in advisory ones (e.g., counselling a friend about a problem). It also showed that socioeconomic status did not predict higher competences in English, although it did so for competences in the school language (German) and mathematics (Hartig & Jude, 2008).

For Switzerland, a recent study in Year 9 in the canton of Aargau showed that 56% of learners had already reached level B1.1 or B1.2 in English writing while 44% were still at levels A1 and A2 (Bayer & Moser, 2016). In an earlier study, Keller-Bolliger (2012) found the same split of writing competence in Year 9 in different regions of Switzerland, with about half the learners still writing at A1/A2 levels and the other half writing at the B1 level.

2.4. Development of English competences at upper secondary level

No empirical studies of English competences of any sort are available for Switzerland beyond Year 9. For Germany, empirical studies at upper secondary level are available only for receptive skills (i.e., reading and listening) and general language proficiency (i.e., C-Tests). These skills play an important role for source-based writing as this requires the synthesis of spoken and printed input under one leading question. Moreover, the effect sizes for reading and listening can serve as a point of comparison for the effect sizes of writing reported in the current study.

Leucht, Fleckenstein, and Köller (2016) showed that, on average, students of the academic track in the German federal state of Schleswig-Holstein reached level B2+ in English listening and reading proficiency at the time of their school-leaving examination in grammar school. In addition, two studies in the federal state of Hamburg measured English skills in the final two years of upper secondary education by means of C-Tests.

In the first study, Trautwein, Köller, Lehmann, and Lüdtke (2007) found an annual gain of $d = 0.29$ while the annual gain reported in the second study was only $d = 0.19$ (Vieluf, Ivanov, & Nikolova, 2014). Leucht, Retelsdorf, Pant, Möller, and Köller (2015) collated data from several studies relating to EFL receptive competences to model learning gains from Year 9–13, again for Schleswig-Holstein. For reading competences, they found a marked increase in proficiency from Year 9 and Year 11 ($d = 0.81$). For Years 11–13, students only showed gains of $d = 0.25$, signifying a marked deceleration of development. As tests had been appropriately designed and validated to measure the full range of proficiency levels effectively, the results can be interpreted as a slowing-down in reading competences due to a saturation effect: realising that their students already had good reading skills, teachers invested less time and energy in developing them and focused on other areas instead.

Overall, the results suggest that the effectiveness of EFL teaching in upper secondary school is limited because learning rates per year correspond to effect sizes that vary between $d = 0.20$ and $d = 0.30$. While these gains might appear quite small, the average annual gain in effect size for nationally normed reading tests for school Years 11–12 reported in a meta-analysis by Bloom, Hill, Black, and Lipsey (2008) was only $d = 0.06$. For subjects such as mathematics, science, and social science the effect sizes were equally small. In sum, the available literature suggests that skills of students in upper secondary education – in foreign languages and other subjects – do not increase substantially over one school year.

2.5. Individual factors predicting EFL (Writing) competence

In this section, we focus on the influence of gender and language background as moderating variables of EFL competence.

2.5.1. Gender

Most studies that are comparable to the current one in terms of geographical and cultural context show higher learning gains in both school language (L1) and second or foreign language (L2) for female in comparison to male students, irrespective of background or school level (Zimmer, Burba, & Rost, 2005). In the Swiss national evaluation of educational standards for foreign languages in compulsory schools for Years 6–9, results indicated that females did significantly better at writing English texts than boys in Year 9, a result which was visible in all populations and subpopulations (e.g., regions, school types) and across social backgrounds (Schneider, Lenz, & Studer, 2009).

Studies in Germany show a similar divide between the genders in the development of English competences. For Year 9, the DESI study showed that female students had significantly higher English competences overall ($d = 0.29$) and that the gender differences were largest for EFL writing skills ($d = 0.51$; Hartig & Jude, 2008). A comparative study of language competences in different German states in Year 9 again showed that female learners had better competences in EFL reading ($d = 0.19$) and EFL listening ($d = 0.16$) than male learners (Winkelmann & Groeneveld, 2010).

2.5.2. Language background

Another significant factor impacting the development of students' competence in EFL writing is language background (i.e., plurilingualism or proficiency in different languages). There are rising numbers of students with an immigration background in

Europe, which means that, for them, the official language of a country is an L2 while English becomes an L3 (Poarch & Bialystok, 2017).

In a re-analysis of data from the DESI study, Göbel, Rauch, and Vieluf (2011) found that students who had grown up using two languages – rather than German alone – showed significantly higher proficiency levels in English reading at the end of lower secondary education. Fleckenstein, Möller, and Baumert (2018) showed that in the context of dual immersion in Germany, bilingual students have an advantage when it comes to learning an additional language.

In a study involving children of Turkish language background in Year 9, however, Rauch, Jurecka, and Hesse (2010) found no effect of bilingualism on English reading competences after controlling for school type and socioeconomic status variables. In a longitudinal study from Years 6–8, Maluch, Neum, ann, and Kempert (2016) found that students with bilingual backgrounds initially showed significant advantages in their acquisition of English competences, but these advantages disappeared over time. A possible explanation for these disparate results is that plurilingualism has different effects in different contexts. It would seem that schools – or society at large – do (too) little to foster the potential inherent in migration-related plurilingualism, especially if the additional languages are not held in high regard in a country (Bild & Swain, 1989).

3. Methods

3.1. Research questions

Informed by the above findings, we explored the following research questions in this study:

(1) Do students' writing competences differ according to task type (argumentative vs. source-based writing)?
(2) What level of overall English writing proficiency do students at German and Swiss Gymnasiums achieve approximately two years, and approximately one year before graduation?
(3) Do English writing competences improve over the course of approximately one school year? (8 months)
(4) What is the influence of the selectivity of school system and the variables "gender" and "language background" on English writing proficiency and its development?

To answer research question 1, we analyze results from the two types of writing tasks used in this study separately. The two genres can be broadly described as argumentative writing (*independent task*) and source-based writing (*integrated task*, see below). We expect students to do slightly better at argumentative than at source-based writing because argumentative writing figures prominently in the EFL curricula of Germany and Switzerland (EDBS, 2017; Standing Conference of the Ministers of Education and Cultural Affairs in the Federal Republic of Germany KMK, 2014). The prompts typically associated with argumentative writing are a staple of upper secondary classrooms (Keller, 2013). We further expect the scores from the integrated task to correlate more highly with the scores from measures of receptive competences because this task requires students to understand and integrate written and auditory stimulus material.

For research question 2, we expect a significant tier of students to reach CEFR level B2 one year ahead of graduation. Only a minority should still be at level A2, as reaching level B2 within a year would be unlikely for that cohort. Furthermore, we expect that only a very slim population of students should perform at the level C2 because reaching near-native writing proficiency solely by formal EFL instruction is improbable. Test scores in this tier might come from students who spent substantial time living abroad or from native-English-speaking students.

Concerning research question 3, we expect a modest increase in writing skills over the period of eight months as EFL writing is an important aspect of educational curricula at that level. Based on previous findings on receptive English skills and C-Tests in upper secondary education, we expect an increase of achievement over eight months corresponding to an effect size of $d = 0.20$ to $d = 0.30$. We are aware that these expected gains are relatively small but they are in line with results from previous studies.

As the Swiss school system at the upper secondary level is more selective than the German one, we expect Swiss students to outperform German students at least at the first timepoint of data collection. Since this can be attributed to higher selectivity, the effect should disappear once results are controlled for background variables.

Concerning research question 4, we expect female students to do slightly better than males in line with international research findings. In terms of language background, two outcomes are conceivable: on the one hand, bilingual students may outperform monolingual ones in English writing because of the potentials of bilingualism for third language learning outlined above. On the other hand, previous research in the German context has shown no consistent advantage for bilingual over monolingual students.

3.2. Sample and procedure

This study was carried out as a repeated measurement design in upper secondary schools in Germany and Switzerland with an interval of eight to nine months between the two measurement points, depending on availability of schools for testing (T1: August/September 2016; T2: May/June 2017). Students completed computer-based tests on writing, reading and listening skills, as well as general cognitive ability. Furthermore, they completed a questionnaire measuring background variables as well as individual characteristics and their perceptions of their English classes in school (Fleckenstein et al., 2019; Rupp et al., 2019).

In both countries, the target population consisted of all students attending the academic track of general education grammar schools (ISCED level 3a). The study focussed on classes that were two years before graduation at the first measurement point and one

year before graduation at the second measurement point. In Switzerland, data were collected from schools in the cantons of Aargau, Basel Stadt, Basel Land, Luzern, St. Gallen, Schwyz, Zurich (all of them have German as L1). There is a slight variation in the length of upper secondary education as some schools take 12 and some 13 years to reach baccalaureate level, resulting from different entry times into the Gymnasium. This resulted in some classes following the 12 year schedule while others followed the 13 year schedule or were composed of students from different school types coming together for their final school years. In Germany, data were collected from upper secondary students in the federal state of Schleswig-Holstein, in which the average level of English competency is close to the national average in Germany (Stanat, Böhme, Schipolowski, & Haag, 2016).

The data collection procedures in the two countries differed: In Switzerland, educational departments were contacted for permission to gather data in schools. Where such permission was received, the research team contacted all relevant public schools in the canton and asked them to participate. Originally, the research team had planned to sample students within schools and classes. However, after feedback from schools, it was decided to recruit whole classes in order to make the organization easier for schools. Thus, the sample in Switzerland was a convenience sample. Furthermore, it was left up to schools to decide how many classes from that year should participate in the study. Where possible, classes were selected which had different "special subjects" (i.e., subjects which receive special attention in the curriculum and are given extra lessons in certain semesters such as modern languages, science, and economics) in order to maximise the representativeness of the sample. This procedure resulted in a sample size of $n = 1882$ from 91 classes nested in 20 schools from the seven cantons.

In Germany, data were gathered in the federal state of Schleswig-Holstein. Data collection was conducted by the International Association for the Evaluation of Educational Achievement Data Processing and Research Center in Hamburg with consent of the Ministry of Education in Schleswig-Holstein. The statistical population consisted of all students attending general education grammar schools with eight years of secondary schooling. A total of 84 out of 99 schools in Schleswig-Holstein followed that model, while 15 schools had nine years of secondary schooling and were thus excluded from the study. For the final sample, 42 schools were randomly selected from the pool of 84 available schools and students were asked for their voluntary participation. In the end, $n = 965$ students from different upper secondary profiles (e.g., language, science, civics) participated. These profiles refer to subjects which are taught at the advanced level at a particular school.

Data collection took place during regular classes in the morning. Intensively trained university student assistants or Ph.D. students supervised the test sessions and instructed all participants. Students always started the test by writing two essays (60 min in total), followed by a test on general cognitive ability (only at T1). After a break, they filled out the questionnaire and worked on listening and reading tests. The overall testing time was three hours at each measurement point. Data was collected on laptops in an offline environment that closely mirrored the Test of English as a Foreign Language (TOEFL) interface (for details of interface design, see Rupp et al., 2019).

Overall, data were collected from $n = 1882$ students in Switzerland (58% female; age: $\bar{x}_{T1} = 17.56$, $SD_{T1} = .91$; $\bar{x}_{T2} = 18.27$, $SD_{T2} = .91$) and $n = 965$ students in Germany (58.6% female; age: $\bar{x}_{T1} = 16.91$, $SD_{T1} = .56$; $\bar{x}_{T2} = 17.61$, $SD_{T2} = .56$). For the analyses, we excluded students who did not take part in the achievement tests at either T1 or T2. In addition, we excluded one school in Switzerland which had not reported the full scope of background information or class membership of participating students. This exclusion of $n = 124$ learners resulted in a final sample of $n = 1729$ students for Switzerland and $n = 894$ students from Germany ($n = 2623$).

3.3. Measures

3.3.1. English writing test

This study aimed for a broad representation of the writing construct represented in the curricula of the two countries. As there was no procedure for large-scale testing of EFL writing skills already established, the research team selected tasks from the TOEFL Internet-based test (iBT) writing test because its writing prompts and assessment rubrics are congruent with the relevant curricula in Germany and Switzerland (Fleckenstein et al., 2019). The conceptualisation of the writing assessment was done in co-operation with the Educational Testing Service (ETS) in Princeton, NJ, USA, which also conducted the scoring of learner texts according to procedures similar to the TOEFL iBT writing assessment (Burstein, Tetreault, & Madnani, 2013; ETS, 2009).

Since it was not possible to use currently operational prompts due to security concerns for the test overall, we opted to use prompts that had been previously designed and were publicly available on the TOEFL website. Out of the pool of all possible tasks, four prompts were deemed best suited in terms of content familiarity and overall difficulty (Table 1). The feasibility of these prompts

Table 1
Task Types and Topics from TOEFL iBT Writing Tasks Used in this Study.

Task Type	*Name*	*Content*
Independent	"Teachers"	A teacher's ability to relate well with students is more important than excellent knowledge of the subject being taught.
	"Advertising"	Television advertising directed toward young children (aged two to five) should not be allowed.
Integrated	"Voting Machines"	A text and lecture on the use of different voting systems used in the United States. Participants were asked to compare traditional voting systems with computerized systems and discuss the value of each system contrastively.
	"The Chevalier"	A text and lecture on the memoirs of the Chevalier de Seingalt (1725–1798, more widely known as Giacomo Casanova). Participants were asked to discuss conflicting statements relating to the reliability of the Chevalier's memoirs, who liked "to make his life seem more exciting and glamorous than it really was."

was tested in several pilot classes in both countries. Even though text quality of the resulting responses from try-outs varied widely across the prompts, there were neither bottom nor ceiling effects. That is, learners seemed able to cope well with the demands in terms of length, level of difficulty and specificity of topic. In addition, the response characteristics fell within the range of abilities recorded in existing TOEFL samples and could be scored by human raters as well as AEE procedures of ETS.

By employing both independent and integrated tasks, this study closely mirrored the writing section of TOEFL iBT. By design, the two task types differ in length and context of writing as well as in the competences required to complete them. The independent task consists of argumentative writing on a controversial topic which is formulated in an *agree-disagree* format (ETS, 2009). Topics are typically chosen in a way that they fall within students' world knowledge, and learners are required to give reasons that support their arguments (i.e., reasons for agreeing or disagreeing with the statement). Learners have 30 min to plan and write a response, explaining their opinion. For our study, we advised learners that a good essay should be at least 300 words long yet told them to not count words. Instead, they were told that 300 words would account for approximately ten lines on the computer screen.

The integrated task requires learners to read a passage that is about 250–300 words long. Afterwards they listen to a lecture of two to three minutes challenging the ideas expressed in the reading passage. Finally, they are given 20 min to plan and compose a short essay of about 150–225 words, summarizing the points made in the lecture and explaining how they challenge the reading passage. This task thus involves "synthesis texts", requiring a combination of different language skills and strategies (van Ockenburg et al., 2016). These include demonstrating an exploratory understanding of the text and lecture given as input material, selecting sensible information from them, and evaluating and arranging it in a logical way (van Ockenburg et al., 2016). Learners are not expected, however, to state their own opinion for this task, but, rather, are asked to synthesize or contrast the information contained in the input material (Plakans, 2010).

Taken together, these two types of writing tasks operationalize two central goals (or functions) of English writing expressed in German and Swiss EFL writing curricula. The first function is for students to be able to state their personal opinion about a given topic, to be persuasive, and to use the foreign language to address a specific readership in an appropriate way. The second function is for students to be able to understand a wide range of spoken and written text types (e.g., reports, lectures) and to either use them as input for their own writing or to transform one genre into another (EDBS, 2017; IQSH, 2014; Fleckenstein et al., 2018).

At each time point, the participants worked on one independent and one integrated prompt and the corresponding complementary prompt for each task type at the other measurement time point in a repeated-measures design. Tasks were permutated in a matrix design to prevent sequence and task effects. A detailed testing manual specified procedures for instructing learners how to take the test, which were kept identical each time.

Text analysis (i.e., human and machine scoring) was done by ETS. All essays were scored by experienced human raters on the operational holistic TOEFL iBT scale from 0 to 5 with each essay receiving two independent ratings. Independent essays were scored high if they were well organized and individual ideas were well developed, if they used specific examples and support to express learners' opinion on the subject, and if English was used accurately to express learners' ideas. Integrated essays were scored high if they clearly presented main points, contrasted lecture and reading points, and contained few grammatical/spelling errors. Essays were assigned a score of 0 if they were written in another language, were generally incomprehensible, or if no text was entered.

The research team hired experienced raters from the operational TOEFL to obtain statistically reliable human ratings that achieved a sufficient degree of construct coverage through the consistent application of the operational TOEFL iBT scoring rubrics. Inter-rater agreement, as measured by quadratic weighted kappa (QWK), was satisfying for both text types over the two time points (QWK = .639/.670 for the two independent prompts and QWK = .865/.775 for the two integrated prompts) (Hayes & Hatch, 1999; see Rupp et al., 2019 for further details).

3.3.2. Automated essay evaluation (AEE)

AEE scoring models for essays involve processing the digitally collected written responses via computational routines, which results in a set of statistical variables called *features* that can be used as predictor variables in statistical models to yield predicted human scores for these essays (Rupp et al., 2019). The development of features is scientifically grounded in the disciplines of natural language processing, computational linguistics, and computer science (Burstein et al., 2013). Arguments for the use of AEE include reliability and ease of scoring as well as the ability to safeguard against human rater errors such as changes over time and topic-, sequence- or leniency effects (Deane, 2013).

While they bolster the objectivity of writing scores in large-scale contexts, AEE systems do not actually "understand" essays as humans do. Whereas human raters directly evaluate various variables of interest such as diction, fluency, and grammar in order to produce an essay score, AEE systems use approximations or possible correlates of these variables (e.g., Attali, 2007; Attali & Burstein, 2005). Since current-generation AEE systems do not model the human ability to perform social and conceptual reasoning, they necessarily measure a narrower range of skills than human raters. Specifically, their features predominantly deal with textual qualities at the level of sentences and phrases, addressing text production abilities that enable writers to organize text according to some outline and elaborate upon the points it contains while using appropriate, clear, concise and unambiguous language in conventional orthography and grammar. There is, however, a deep connection between the linguistic skills measured by AEE and the broader bundle of abilities included in human ratings: students who have mastered "basic" L2 skills such as vocabulary, grammar and mechanics also have more cognitive resources to master a broader writing construct (Deane, 2013).

For our study, each text was scored by *e-rater®*, the operational AEE engine of the TOEFL iBT test (Burstein et al., 2013). The *macrofeatures* (or main features) used to score responses were grammar, usage, mechanics, organization, development, discourse, collocations and prepositions, average word length, median word frequency, and sentence variety (Rupp et al., 2019). As the current study also used prompt-specific scoring models, the model included prompt-specific vocabulary usage measures (Attali, 2007). In

addition, we used two independent blind human ratings so that human ratings received twice the weight of automated scores.

3.3.3. Background variables

Information on students' gender was collected from the schools' participation lists. Questionnaires were used to determine language background and socioeconomic status (SES). In order to assess language background students were asked which languages they spoke at home. If students indicated two or more home languages, they were classified as bilingual; if they only indicated one language, they were classified as monolingual.

We measured SES using the Highest International Standard Classification of Education (HISCED) proposed by the United Nations Educational, Scientific and Cultural Organization (UNESCO, 2011). HISCED is coded as the highest vocational education certificate obtained by one of the parents in a family (classified by each parent's ISCED). General cognitive ability was assessed using the subtests on figural reasoning (N2) and on verbal reasoning (V3) of the cognitive ability test (Heller & Perleth, 2000).

3.3.4. Proficiency measures

In order to obtain reliable and non-redundant proficiency scores for writing, we used common item response theory (IRT) scaling techniques, which involved computing expected a posteriori (EAP) estimates of scores. Since the study also tested learners' receptive skills, the scaling was done by first estimating item parameters and EAP person parameters from a three-dimensional longitudinal model (reading, listening, writing) for each learner using IRT scaling techniques as implemented in M*plus* Version 8 (Muthén & Muthén, 1998–2017). To determine whether student performance differed by task type or genre, we also computed separate unscaled scores (i.e., averages of two human and one machine [hhm] score) for independent and integrated prompts. To investigate the effect of task type, we conducted a multivariate analysis of variance (MANOVA) for the raw independent and integrated scores with time point and task type as within-subject factors and country as a between-subject factor.

The listening and reading skills were measured with a subset of items from the German National Assessment. Those tests consist of 71 tasks with 391 items for reading, and 65 tasks with 352 items for listening, testing the entire range of competence levels of the CEFR (A1 to C2) on one integrated competence model (Stanat et al., 2016). The items used in our study required the detailed understanding of long, complex reading and listening texts including idiomatic expressions and different linguistic registers.

Two hhm scores for two essays per point of measurement served as indicators for writing ability; EAP reliabilities for all three domains were above .75. Together with a comprehensive background we estimated 15 plausible values (PV) for each domain per student and point of measurement. The PVs were then standardised and transformed to a metric with $M = 500$ and $SD = 100$ at T1. By standardising all values along the values for T1, differences between T1 and T2 can be directly interpreted as changes between the measurement points.

3.3.5. CEFR classifications

In order to report proficiency levels of students' writing we established a link of the TOEFL iBT scores to CEFR levels. To calculate this measure, we scaled both task types jointly to create a general indicator of writing competence and linked the results to the CEFR. Specifically, we conducted a *standard-setting* study (Fleckenstein et al., 2019) using a modified version of the *performance profile method* (Hambleton, Jaeger, Plake, & Mills, 2000), which yielded cut-scores that we used to relate the TOEFL iBT scores to the CEFR levels. The score ranges (transformed to $M = 500$; $SD = 100$) for the relevant CEFR levels were as follows: A2 (< 419); B1 (419–483); B2 (484–613); C1 (613–741); C2 (> 741).

3.3.6. Statistical analyses

All analyses were conducted in M*plus* version 8 (Muthén & Muthén, 1998–2012) based on the 15 PV data sets using robust maximum likelihood estimation to account for a hierarchical data structure (i.e., students clustered in classes; type = complex). Full-information maximum likelihood was used to estimate missing values in background variables. Due to the use of 15 PVs, all analyses were run 15 times and then averaged (see Rubin, 1987). The expected differences in writing scores between countries were investigated with linear regressions models.

4. Results

4.1. General validity measures for writing scores

Human-machine agreement was satisfactory for all writing tasks as score correlations ranged from $r = .762$ to $r = .807$ for the independent prompts and from $r = .715$ to $r = .814$ for the integrated prompts. This suggests that custom-built, prompt-specific AEE models created performed satisfactorily for our mixed population and tasks (Rupp et al., 2019). The scores for the four prompts were also moderately correlated at $r = .51$ to $r = .60$ in the pooled sample.

We also computed correlations between writing scores and secondary measures (listening, reading and various school grades). The writing scores for all four prompts were positively correlated with the scores from the listening and reading comprehension assessments, with correlations ranging from $r = .554$ to $r = .725$ for listening and from $r = .415$ to $r = .592$ for reading, and similar patterns for both countries. As expected, correlations of writing scores with scores from measures of receptive skills were slightly higher for the integrated than for the independent tasks. For example, scores from reading measures at T2 in the Swiss sample correlated with scores from the independent tasks at $r = .576$ (Teachers) and $r = .563$ (TV Advertising) and with scores from the integrated tasks at $r = .694$ (Chevalier) and $r = .640$ (Voting Machines) (Rupp et al., 2019). While these correlations between

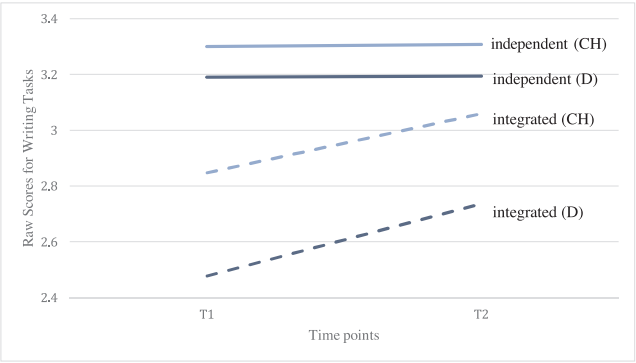

Fig. 1. Performance on integrated and independent Tasks at T1 and T2 (averaged hhm scores).

receptive and productive competences are larger than the ones reported in the DESI study in Year 9 (.28–.41; Hartig & Jude, 2018), the differences in correlational patterns between independent and integrated task types were rather small.

Significant correlations were also found between school grades in English and all four essay scores across samples, ranging from $r = .373$ to $r = .557$ (convergent validity). Lower correlations between writing scores and school grades were found for German, ranging from $r = .113$ to $r = .400$, while the smallest correlations were found for mathematics, which ranged from $r = -.012$ to $r = .293$ (discriminant validity).

4.2. Research question 1

As Fig. 1 reveals, students in both countries achieved significantly higher average scores on the independent tasks than on the integrated tasks (main effect task type: $F(1, 1977) = 96.704$, $p \leq .001$). This is likely due to the higher familiarity with the independent than the integrated task type. Both tasks showed an increase between the two time points (main effect time: $F(1, 1977) = 1315.338$, $p \leq .001$). However, Fig. 1 also reveals that the increase is larger for the integrated tasks, while there is only a slight increase for the independent tasks. This is confirmed by the significant interaction term between the measuring point and the task type in the MANOVA ($F(1, 1977) = 98.845$, $p \leq .001$). One explanation would be that learners already did well on the independent tasks at T1, creating a ceiling effect for the independent task. The same was not true for the integrated tasks, however, leaving more room for improvement at T2.

As expected, Swiss students outperformed their German counterparts in both task types, which can be attributed to the higher selectivity of the Swiss system (main effect country: $F(1, 1977) = 64.112$, $p \leq .001$). The MANOVA also shows an interaction effect between country and time point (interaction effect country/time point: $F(1, 1977) = 82.488$, $p \leq .001$), which might be due to a slightly higher improvement in integrated task performance in Germany (GE: $\text{diff}_{T1\text{-}T2} = .259$ vs CH: $\text{diff}_{T1\text{-}T2} = .213$), resulting in a slightly steeper increase.

4.3. Research question 2

The second research question related to students' proficiency levels in EFL writing at the beginning and end of their penultimate year before graduation. Recall that educational standards in both countries mandate that students reach the target of CEFR level B2 at the end of upper secondary education.

Results show that the majority of students in our study already reach level B2 at T1, with only 35.3% of students failing to reach that minimal standards for upper secondary education. That percentage decreased to 28.2% at T2. This is a positive result as it shows some progress over eight months, yet it also implies that nearly 80% of the students that had not achieved B2 early on did not achieve it at all in secondary education. A substantial number of students reached level C1 (effective operational proficiency) at T2, yet only a handful reached the highest CEFR level C2 (mastery).

Due to the higher selectivity of Swiss upper secondary schools, we expected Swiss students to outperform their German counterparts. The larger group of students that reached B2 and C1 in Switzerland compared to Germany seems to confirm this expectation.

4.4. Research question 3

The third research question relates to the development of writing competences between T1 and T2. As Fig. 2 shows, the percentage of students writing on the B2 and C1 level increased from T1 to T2, suggesting a moderate development of proficiency over time. Table 2 shows that this increase is indeed significant (column MT2 - MT1) independent of country, gender, and language background. This conforms to expectations as English writing figures prominently both in Germany and Swiss curricula for that school year. Effect sizes for the development of English writing proficiency were small but in line with expectations derived from studies in similar contexts ($d = 0.15$–0.20).

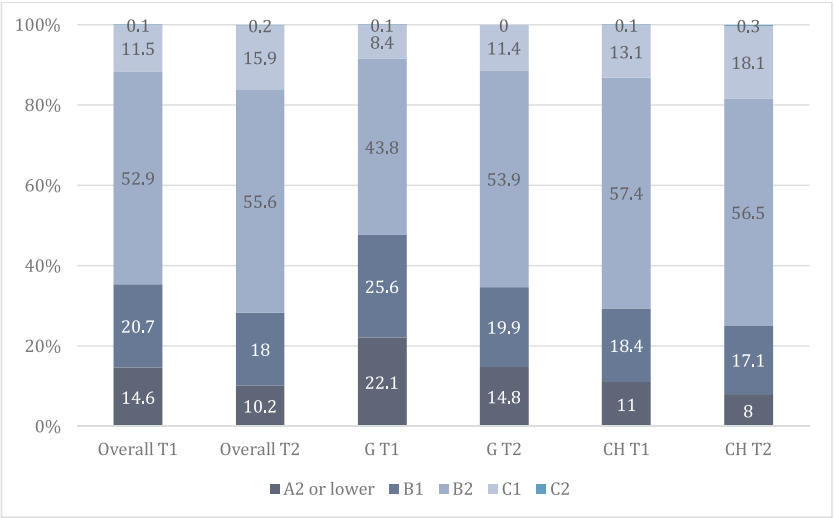

Fig. 2. CEFR levels for English Writing at T1 and T2 in Germany (G) and Switzerland (CH).

Table 2
Writing Competences (Means and Standard Deviations) at T1 and T2 by Country, Gender, and Language Background.

	Writing					
	T1		T2		Delta	
	M	*SD*	*M*	*SD*	$M_{T2} - M_{T1}$[1]	*D*
Overall (N = 2623)	500.00	99.98	518.65	107.37	18.65**	0.18
Country						
Germany (n = 894)	472.419	103.18	493.166	110.17	20.747**	0.19
Switzerland (n = 1729)	514.261	95.20	531.832	103.45	17.571**	0.18
Gender						
Female (n = 1535)	500.380	97.61	520.620	105.82	20.24**	0.20
Male (n = 1088)	499.464	103.23	515.879	109.44	16.41**	0.15
Language background[2]						
Monolingual (n = 1821)	497.86	99.84	513.91	106.32	16.05**	0.16
Bilingual (n = 469)	504.86	93.64	522.72	106.62	17.87**	0.18

[1] Wald tests in M*plus* 8 were used to estimate parameter differences.
[2] Students with English as a home language (n = 136) were excluded from the analyses for the language background.
** $p < .01$. The proficiency scores metrics were set to M = 500 and SD = 100 at T1.

Table 3
Findings of Linear Regression Analysis Relating English Writing Skills at T2 with Background Variables and Writing Skills at T1.

Intercept/predictors/R^2	Model 1	Model 2	Model 3
Intercept	490.32 (6.57)**	491.845 (6.421)**	510.685 (3.801)**
Country (0 = Germany; 1 = Switzerland)	38.68 (7.58)**	26.279 (7.405)**	4.469 (4.284)
Gender (0 = male; 1 = female)	4.85 (4.82)	7.352 (4.803)	5.006 (3.505)
General cognitive ability (KFT; z-standardized)		25.272 (2.662)**	5.507 (2.037)**
Socioeconomic status (HISCED; z-standardized)		6.403 (2.508)*	1.301 (2.036)
Language background (0 = monolingual; 1 = bilingual)		10.356 (6.907)	4.449 (4.975)
Writing T1			74.526 (2.438)**
R^2	0.03 (0.01)**	0.09 (0.01)**	0.52 (0.02)**

Note. Unstandardized Regression Weights and Standard Errors (in Parentheses). Analyses are based on a sample of n = 2487 excluding those students that speak English as a home language. The metrics for the proficiency variables were set to M = 500 and SD = 100 at T1. All coefficients are unstandandardized. *$p < .05$, **$p < .01$.

4.5. Research question 4

The fourth research question concerned the effects of selected background variables and institutional differences across countries on writing competence; results are shown in Table 3.

Findings show that the average score difference between Swiss and German students was indeed significant, as Swiss students outperformed German students by more than a third of a standard deviation (Model 1). Note that eight months of English learning corresponds to an increase of 20 points on our scale ($M = 500$; $SD = 100$), suggesting that Swiss students were about 15–16 months ahead of German students. The effects remained substantial even after including cognitive, socioeconomic and language background factors (Model 2). The advantage of Swiss students disappears, however, once T1 writing performance is entered into the model (Model 3). This indicates that students in both Germany and Switzerland have about the same learning gains over the course of 8 months but started from different baselines at the beginning of the school year; Fig. 2 (above) supports this assumption.

Contrary to our expectations, gender did not have a significant effect on writing competence in either of the models. Similarly, students' language background influenced neither the proficiency level nor the development of writing competences. General cognitive ability had a significant and substantial effect on T2 writing skills. The regression coefficient for this variable in Model 2 (Table 3) indicates that an increase of one standard deviation in cognitive abilities resulted in an increase of the writing scores of 25 points, which corresponds to one year of schooling. The effect of cognitive abilities on writing skills was quite small but significant, even after controlling for T1 writing skills. Finally, the effect of socioeconomic status was significant (M2) but rather small and disappeared after controlling for T1 writing skills (M3).

5. Discussion

The aim of this study was to measure writing competences of learners in the academic track of general education grammar schools in Switzerland and Germany. It is the first representative study of EFL writing beyond Year 9 in either country. It is relevant beyond this context, however, because argumentative and integrated writing are key foreign-language competences of upper-secondary education in many countries world-wide. Further, CEFR level B2 is generally accepted as entry level into tertiary education in the entire European Union (European Commission, 2017). Results suggest that, on average, learners show acceptable levels of proficiency in the penultimate year of their baccalaureate education. On average, they were able to write English essays that address real-world, controversial issues and use somewhat developed explanations, exemplifications, and details of argumentation to support their points of view. Their texts displayed sufficient unity, progression, and coherence for readers to understand the main message, though connection of ideas was occasionally obscured.

On the basis of the results presented in this paper, it seems that B2 is a realistic target level for upper secondary schools, both in participating countries and other European nations. There is a small but significant tier of students already writing at the C1 level of CEFR who are especially well equipped for the challenges of tertiary education. Our results also show, however, that students who did not achieve B2 early on tended not to achieve it at all, suggesting a lack of progress in the group of students who needs it most. This should be explored in further studies.

At the level of task types, students seem to do better at argumentative than at source-based writing, which we attribute to a lack of task familiarity for the latter type of task. This would imply that writing synthesis texts (van Ockenburg et al., 2016) should become an explicit focus of teaching at upper secondary level, as they are essential for tertiary education. However, scores from both task types showed similar correlations with scores from measures of receptive competences, suggesting that performance in the integrated task is not determined by students' reading or listening skills alone. Rather, the results could be interpreted along the lines of Schoonen's (2019) hypothesis that declarative and metacognitive knowledge of L2 use provide a rich source from which both receptive and productive competences draw.

These results have implications both for teacher education and classroom practice. First, teacher education should focus on familiarising teachers more deeply with the relevant genres for tertiary education. Teachers need to become more familiar with the internal workings of key genres to support learners in mastering them. Source-based writing – as operationalised by the integrated tasks in this study – seems especially in need of attention. This includes a focus on teachers' diagnostic competences to identify students who are not achieving the required levels of competence.

Analyses involving control variables showed no significant influence of either gender or language background on the development of writing skills. We take this as tentative evidence that schools manage to implement educational programs which are appealing and motivating to both genders and students with different language backgrounds. In line with other international studies (Programme for International Student Assessment/PISA, 2000; 2003), we found that both cognitive abilities and SES were a significant influence on learners' school achievements. However, the effect of cognitive abilities was much larger compared to SES, suggesting that cognitive entry characteristics are more important for achievement outcomes in upper secondary school than SES. This finding is in line with the literature cited at the outset of this paper, which underlines that, after controlling for prior achievement and cognitive abilities, effects of social background mainly disappear (Hartig & Jude, 2008). Therefore, there is not much evidence for negative correlational effects of SES on writing skills in upper secondary education.

Finally, achievement gains over a time period of eight months were relatively small ($d \sim 0.20$), which corresponds to previous research on English receptive skills and on C-Test performance (e.g., Vieluf et al., 2014). Gains reported in this study are slightly larger than the annual gains typically found in other subjects in upper secondary school (Bloom, Hill, Black, & Lipsey, 2008). However, their magnitude may be disappointing for practitioners who strive to increase their students' skills.

5.1. Limitations

The fact that, on average, students reached level B2 one year before graduation suggests that they have a solid basis for the challenges of tertiary education. Yet the writing measures used in this study do not permit a more detailed look at the internal structure of students' writing skills like their ability to use a thesis statement or to structure an argumentative essay effectively (Brupbacher et al., 2008). While the holistic TOEFL scoring rubrics used in this study do contain references to texts effectively addressing a topic or displaying unity of progression (ETS, 2009), no separate ratings were obtained for aspects such as force of argumentation, appropriateness of examples, and so on. Therefore, a detailed linguistic analysis of prototypical learner essays at different ability ranges is currently being undertaken in a dissertation to get a more fine-grained picture of students' writing competences.

Our study was further limited in terms of its sample, which included only the highest achieving tier of students in both countries. To get a more comprehensive picture, the scope of future studies should be broadened to other tracks of upper secondary education (e.g., vocational schools), as strong English writing competences are necessary for those students as well. Based on studies on English receptive competences in different upper secondary school types (Leucht et al., 2015), it stands to reason that learners in vocational schools do not reach the relatively high standards of students in the academic track.

Moreover, even though the kind of writing that is represented by the two TOEFL tasks is in line with expectations about writing proficiency as articulated in standards such as the CEFR, it is unclear how much time, exactly, teachers in the different classrooms spent on teaching this kind of writing. This includes uncertainty about how much time they spent on providing truly diagnostic feedback based on qualitative analyses of submitted responses outside of this study. Similarly, while these tasks are highly relevant in terms of the curricular alignment, they are isolated assessment prompts in which responses are timed, collaboration with others is not possible, and creating multiple drafts is not advisable. Further studies relating to more authentic tasks should follow. In particular, we need more insight into students' processes of learning to write in response to specific prompts, materials and teacher input. This research team is planning to conduct such studies in the future.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This study is supported by the Swiss National Science Foundation (SNF; Grant No. 100019L162675) and the German National Science Foundation (DFG, Grant No. KO1513/12-1).

References

Attali, Y. (2007). *Construct validity of e-rater in scoring TOEFL essays. ETS research report RR-07-21* Retrieved from https://www.ets.org/Media/Research/pdf/RR-07-21.pdf.

Attali, Y., & Burstein, J. (2005). *Automated essay scoring with e-rater® V.2. (ETS research report RR-04-45)* Retrieved from https://www.ets.org/Media/Research/pdf/RR-04-45.pdf.

Bayer, N., & Moser, U. (2016). *Evaluation der Englischkompetenzen im Kanton Aargau. Englischkompetenzen auf der Primarstufe und auf der Sekundarstufe I [Evaluation of English competences at primary and secondary schools in the canton of Aarau]*. Retrieved from https://www.ibe.uzh.ch/de/publikationen.html.

Bild, E. R., & Swain, M. (1989). Minority language students in a French immersion programme: Their French proficiency. *Journal of Multilingual and Multicultural Development, 10*(3), 255–274. https://doi.org/10.1080/01434632.1989.9994377.

Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness, 1*(4), 289–328. https://doi.org/10.1080/19345740802400072.

Brupbacher, B., Jucker, A., König, E., Roth, M., & Straumann, B. (2008). Englisch. In Hochschule und Gymnasium [University and Gymansium; HSGYM] (Ed.). *Hochschule und Studierfähigkeit [University and the ability to study]* (pp. 89–96). Retrieved from https://www.hsgym.ch/.

Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater® automated essay scoring system. In M. D. Shermis, & J. Burstein (Eds.). *Handbook of automated essay evaluation: Current applications and new directions* (pp. 55–67). New York, NY: Routledge.

Council of Europe (CoE) (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Cambridge, U.K: Cambridge University Press.

Council of Europe (2008). *Recommendation 2008/7*. Retrieved from https://www.coe.int/en/web/common-european-framework-reference-languages/extracts-recommendation-2008-7.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing, 18*(1), 7–24. https://doi.org/10.1016/j.asw.2012.10.002.

Educational Department of Basel-Stadt [EDBS] (2017). *Lehrplan Gymnasium [Curriculum for Gymnasium schools]*. Retrieved from www.edubs.ch.

Educational Department of Berne [EDB] (2017). *Sprachniveau an der Maturität gemäss Europäischem Sprachenportfolio (ESP) [Language proficiency at Matura level according to the European Language Portfolio ELP]*. Retrieved from https://www.erz.be.ch/erz/de/index/mittelschule/mittelschule/publikationen.assetref/dam/documents/ERZ/MBA/de/AMS/amssprachniveaumaturitaet.pdf.

Educational Testing Service [ETS] (2009). *The official guide to the TOEFL test.* New York: McGraw-Hill.

European Commission/EACEA/Eurydice (2017). *Key data on teaching languages at school in Europe – 2017 edition. Eurydice Report* Retrieved from Publications Office of the European Union website: https://publications.europa.eu/en/publication-detail/-/publication/73ac5ebd-473e-11e7-aea8-01aa75ed71a1/language-en/format-PDF.

Fleckenstein, J., Keller, S., Krüger, M., Tannenbaum, R., & Köller, O. (2019). Linking TOEFL iBT® writing rubrics to CEFR levels: Cut scores and validity evidence from a standard setting study. *Assessing Writing*. https://doi.org/10.1016/j.asw.2019.100420.

Fleckenstein, J., Möller, J., & Baumert, J. (2018). Mehrsprachigkeit als Ressource: Kompetenzen dual-immersiv unterrichteter Schülerinnen und Schüler in der Drittsprache Englisch [Multilingualism as a resource: competences of dual-immersive learners in English as a third language]. *Zeitschrift für Erziehungswissenschaft, 21*(1), 97–120. https://doi.org/10.1007/s11618-017-0792-9.

Göbel, K., Rauch, D., & Vieluf, S. (2011). Leistungsbedingungen und Leistungsergebnisse von Schülerinnen und Schülern türkischer, russischer und polnischer Herkunftssprachen [Determinants and results of learning outcomes by students with Turkish, Polish and Russian as first language]. *Zeitschrift für Interkulturellen Fremdsprachenunterricht, 16*(2), 50–65 Retrieved from https://tujournals.ulb.tu-darmstadt.de/index.php/zif/article/view/118/113.

Hambleton, R. H., Jaeger, R. M., Plake, B. S., & Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement, 24*(4), 355–366. https://doi.org/10.1177/01466210022031804.

Harsch, C., Schröder, K., & Neumann, A. (2008). Schreiben Englisch [Writing in English]. In DESI-Konsortium (Ed.). *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie [Teaching and competence development in German and English. Results from the DESI-study]* (pp. 139–148). Weinheim: Beltz.

Hartig, J., & Jude, N. (2008). Sprachkompetenzen von Mädchen und Jungen [Language competencies of males and females]. In DESI-Konsortium (Ed.). *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie [Teaching and competence development in German and English. Results from the DESI-study]* (pp. 204–208). Weinheim: Beltz.

Hayes, J. R., & Hatch, J. A. (1999). Issues in measuring reliability: Correlation versus percentage of agreement. *Written Communication, 16*(3), 354–367. https://doi.org/10.1177/0741088399016003004.

Heller, K. A., & Perleth, C. (2000). *KFT 4-12+ R. Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision [Cognitive ability test for class level 4 to 12, revised version].* Göttingen: Beltz.

Institute for Quality Development in Schools of Schleswig Holstein [IQSH] (2014). *Fachanforderungen Englisch Sek I / II [Englisch curriculum secondary level I / II].* Retrieved from https://faecher.lernnetz.de/faecherportal/index.php?key=2&wahl=10436&auswahl=101.

Keller, S. (2013). *Integrative Schreibdidaktik Englisch für die Sekundarstufe. [Integrated writing curriculum at secondary level].* Tübingen: Narr.

Keller-Bolliger, R. (2012). *Kommunikative Schreibkompetenz in der Fremdsprache erfassen und beurteilen [Measuring and assessing communicative writing competences in foreign languages].* Berlin: epubli.

Klieme, E.,, Helmke, A.,, Lehmann, R.,, Nold, G.,, Rolff, H.,, & Schröder, K., (Eds.). (2008). *Unterricht und Kompetenzerwerb in Deutsch und Englisch [Teaching and competence development in German and English].* Weinheim: Beltz.

Leucht, M., Fleckenstein, J., & Köller, O. (2016). Erreichen kriterialer Leistungsstandards in der ersten Fremdsprache Englisch [Learners' achievement of educational standards in English as second language]. In M. Leucht, N. Kampa, & O. Köller (Eds.). *Fachleistungen beim Abitur. Vergleich allgemeinbildender und beruflicher Gymnasien in Schleswig-Holstein [Educational outcomes at Abitur. Comparing academic and vocational Gymnasium]* (pp. 171–199). Münster: Waxmann.

Leucht, M., Retelsdorf, J., Pant, H. A., Möller, J., & Köller, O. (2015). Effekte der Gymnasialprofilzugehörigkeit auf Leistungsentwicklungen im Fach Englisch [Effects of Gymnasium profile on performance development in English]. *Zeitschrift für Pädagogische Psychologie, 29*(2), 77–88. https://doi.org/10.1024/1010-0652/a000153.

Maluch, J. T., Neumann, M., & Kempert, S. (2016). Bilingualism as a resource for foreign language learning of language minority students? Empirical evidence from a longitudinal study during primary and secondary school in Germany. *Learning and Individual Differences, 51*, 111–118. https://doi.org/10.1016/j.lindif.2016.09.001.

Muthén, B. O., & Muthén, L. K. (1998–2017). M*plus* (Version 8) [Computer software]. Los Angeles.

National Statistical Office of Switzerland [Bundesamt für Statistik/BFS] (2016). *Maturitätsquote. [Rate of Matura exams].* Retrieved from https://www.bfs.admin.ch/bfs/de/home.

Plakans, L. (2010). Independent vs. integrated writing tasks: A comparison of task representation. *TESOL Quarterly, 44*(1), 185–194. https://doi.org/10.5054/tq.2010.215251.

Poarch, G. J., & Bialystok, E. (2017). Assessing the implications of migrant multilingualism for language education. *Zeitschrift für Erziehungswissenschaft, 20*(175), 175–191. https://doi.org/10.1007/s11618-017-0739-1.

Programme for International Student Assessment [PISA] (2000). *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich [Basic competences of students in international comparison].* Opladen: Leske & Budrich.

Programme for International Student Assessment [PISA] (2003). *Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs [Level of education of students in Germany – results from the second international comparison].* Münster: Waxmann.

Rauch, D. P., Jurecka, A., & Hesse, H. G. (2010). Für den Drittspracherwerb zählt auch die Lesekompetenz in der Herkunftssprache: Untersuchung der Türkisch-, Deutsch- und Englisch-Lesekompetenz bei Deutsch-Türkisch bilingualen Schülern [First language is key for the acquisition of a third language: A study of reading competences in Turkish, German and English in German-Turkish bilingual students]. *Zeitschrift für Pädagogik, 56*, 78–100 Retrieved from https://www.pedocs.de/volltexte/2012/6946/pdf/Rauch_Jurecka_Hesse_Drittspracherwerb_Lesekompetenz.pdf.

Rijlaarsdam, G., Janssen, T., Braaksma, M., van Steendam, E., van den Branden, K., Couzijn, M., & Verheyden, L. (2013). Learning and instruction in writing. In C. A. Stone, (Ed.). *Handbook of language and literacy: Development and disorders* (pp. 545–566). (2nd ed.). New York: Guilford Press.

Rubin, D. B. (1987). *Multiple imputations for non-response in surveys.* New York: Wileyhttps://doi.org/10.1002/9780470316696.fmatter Retrieved from Wiley Online Library:.

Rupp, A., Casabianca, J., Krüger, M., Keller, S., & Köller, O. (2019). *Automated essay scoring at scale: A case study in Switzerland and Germany (ETS Research Report No. RR-19-12)*Retrieved from Wiley Online Library:https://doi.org/10.1002/ets2.12249.

Rychen, D., & Salganik, L. H. (2003). *Key competencies for a successful life and a well-functioning society.* Göttingen: Hogrefe & Huber.

Schneider, G., Lenz, P., & Studer, T. (2009). *Fremdsprachen – Wissenschaftlicher Kurzbericht und Kompetenzmodell [Foreign languages - scientific report and model of competence]* Retrieved from https://edudoc.ch/record/87025/files/L2_wissB_25_1_10_d.pdf.

Schoonen, R. (2019). Are reading and writing building on the same skills? The relationship between reading and writing in L1 and EFL. *Reading and Writing, 32*, 511–535. https://doi.org/10.1007/s11145-018-9874-1.

Schoonen, R., van Gelderen, A., Stoel, R., Hulsijn, J., & Glopper, K. (2010). Modeling the development of L1 and EFL writing proficiency of secondary school students. *Language Learning, 61*(1), 31–79. https://doi.org/10.1111/j.1467-9922.2010.00590.x.

Stanat, P., Böhme, K., Schipolowski, S., & Haag, N. (2016). *IQB-Bildungstrend 2015. Sprachliche Kompetenzen am Ende der 9. Jahrgangsstufe im zweiten Ländervergleich [IQB educational trend 2015. Language competencies at the end of 9th grade in the second international study].* Münster: Waxmann.

Standing Conference of the Ministers of Education and Cultural Affairs in the Federal Republic of Germany [KMK] (2014). *Bildungsstandards für die fortgeführte Fremdsprache (Englisch/Französisch) für die Allgemeine Hochschulreife [Educational standards for foreign languages (English/French) for general matriculation standard].* Köln: Wolters.

Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany [KMK] (2018). *Schüler, Klassen, Lehrer und Absolventen der Schulen 2007-2016 [Students, classes, teachers and graduates of schools 2007-2016].* Retrieved from https://www.kmk.org/dokumentation-statistik/statistik/schulstatistik/schueler-klassen-lehrer-und-absolventen.html.

Trautwein, U., Köller, O., Lehmann, R., & Lüdtke, O. (2007). *Schulleistungen von Abiturienten [Educational results of upper secondary education graduates].* Münster: Waxmann.

United Nations Educational, Scientific and Cultural Organization [UNESCO] (2011). *International standard classification of educations ISCED 2011.* Retrieved from www.unesco.org.

van Ockenburg, L., van Weijen, D., & Rijlaarsdam, G. (2016). Learning to write synthesis texts: A review of intervention studies. *Journal of Writing Research, 10*(3), 402–428. https://doi.org/10.17239/jowr-2019.10.03.01.

Vieluf, U., Ivanov, S., & Nikolova, R. (2014). *Kompetenzen und Einstellungen von Schülerinnen und Schülern an Hamburger Schulen am Ende der gymnasialen Oberstufe [Competences and Beliefs of Students in Hamburg Schools at the End of Upper Secondary Education].* Retrieved from https://bildungsserver.hamburg.de/contentblob/4396048/6b49c68061321ae400aaa4f7250ebe9f/data/kess12-13.pdf.

Winkelmann, H., & Groeneveld, I. (2010). Geschlechterdisparitäten [Gender differences]. In O. Köller, M. Knigge, & B. Tesch (Eds.). *Sprachliche Kompetenzen im Ländervergleich [Comparison of language competences in German states]* (pp. 177–185). Münster: Waxmann.

Zimmer, K., Burba, D., & Rost, J. (2005). Kompetenzen von Jungen und Mädchen [Competences of males and females]. In PISA Konsortium Deutschland (Ed.). *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs [Level of education of students in Germany – results from the*

second international comparison] (pp. 211–233). Münster: Waxmann.

Stefan D. Keller (Prof. Dr.) is professor of English Teaching and Learning at the School of Education, University of Applied Sciences and Arts Northwestern Switzerland (Institute of Secondary Education). He is deputy director of the Institute for Educational Sciences, University of Basel. Personal information and author photograph: https://www.fhnw.ch/de/personen/stefan-keller

Johanna Fleckenstein (Dr.) is post-doctoral researcher in the Department of Educational Sciences and Psychology at Leibniz Institute for Science and Mathematics Education (Kiel, Germany). Personal information and author photograph: https://www.ipn.uni-kiel.de/de/das-ipn/abteilungen/erziehungswissenschaft/mitarbeiter/fleckenstein-johanna

Maleika Krüger (M.Sc.) is research assistant and doctoral student at the School of Education, University of Applied Sciences and Arts Northwestern Switzerland and at the Institute for Educational Sciences, University of Basel. Personal information and author photograph: https://www.fhnw.ch/de/personen/maleika-krueger

Olaf Köller (Prof. Dr.) is Director and Professor of Educational Psychology at Leibniz Institute for Science and Mathematics Education (Kiel, Germany). Personal information and author photograph: https://www.ipn.uni-kiel.de/de/das-ipn/abteilungen/erziehungswissenschaft/mitarbeiter/koeller-olaf

André A. Rupp (Dr.) is Research Director at Educational Testing Service (ETS) in Princeton, NJ, where he works with teams that conduct comprehensive evaluation work for mature and emerging automated scoring systems. Personal information and author photograph: https://scholar.google.nl/citations?user=RLrt5vMAAAAJ&hl=en