Attila M. Wind* and Anna Zólyomi

The longitudinal development of self-assessment and academic writing: an advanced writing programme

https://doi.org/10.1515/cercles-2022-2046
Received March 10, 2021; accepted November 29, 2021

Abstract: Although several studies have investigated the self-assessment (SA) of writing skills, most research has adopted a cross-sectional research design. Consequently, our knowledge about the longitudinal development of SA is limited. This study investigated whether SA instruction leads to improvement in SA accuracy and in second language (L2) writing. A total of 33 English as a foreign language (EFL) students composed and self-assessed two argumentative essays, one at the beginning (Time 1) and one at the end (Time 2) of a semester-long advanced writing (AW) programme at a Hungarian university. About half of the participants received SA instruction (experimental group), while the other half did not (control group). The essays were scored by two teachers and analysed for linguistic complexity. The results showed improvement in SA accuracy in both groups. However, the SA-teacher assessment (TA) correlation for the total score was statistically significant only in the experimental group at Time 2 (post-instructional phase). Furthermore, the TA total scores and a few linguistic complexity indices showed improvements in L2 writing in both groups. The pedagogical implications of these findings emphasising the importance of SA in EFL writing courses are also discussed.

Keywords: advanced writing programme; linguistic complexity; second language writing; self-assessment; teacher assessment

1 Introduction

Teachers of English as a foreign language (EFL) are required to regularly assess their students' knowledge and performance, according to the National Core

***Corresponding author: Attila M. Wind**, Eötvös Loránd University, Budapest, Hungary, E-mail: wind.attila@btk.elte.hu. https://orcid.org/0000-0002-2702-5211
Anna Zólyomi, Eötvös Loránd University, Budapest, Hungary, E-mail: zolyomi.anna@btk.elte.hu. https://orcid.org/0000-0002-9280-5775

Curriculum of Hungary. Therefore, being able to assess language skills is undoubtedly an important skill for teacher trainees (Hubai and Lázár 2018). However, at universities, language teacher trainees are rarely taught how to assess language skills, especially writing abilities (Csépes 2016). Therefore, the implementation of writing assessment, such as self-assessment (SA) of writing abilities, might be beneficial for future language teachers in the Hungarian educational context. The benefits of the implementation of writing assessment and self-assessment are twofold at the tertiary level. First, teacher trainees will be instructed how to assess writing (e.g., how to design a rubric). Second, they will be made aware of the benefits of self-assessment.

SA is an "internal" approach to measuring language proficiency (Oscarson 1989: 1) which has gained popularity thanks to the increasing interest in learner autonomy and the conceptual change from teacher- to learner-centred instruction (Butler and Lee 2010; Dann 2002). SA facilitates learners' decision-making regarding their language abilities and setting their own goals in language learning (Chapelle and Brindley 2010; Chen 2008). Additionally, as an instrument to explore and understand language performance, SA has also been adopted by the Common European Framework of Reference for Languages (CEFR; Council of Europe 2001), the European Language Portfolio and the Bergen "Can-Do" project (Hasselgreen 2000). National policies have also promoted the implementation of SA in classrooms in Japan and Korea (Butler 2018; Butler and Lee 2006). However, little is known about the implementation of SA in classroom settings in the Hungarian educational context.

The present study attempts to explore the SA of writing abilities of English majors at a university in Hungary. The findings might justify the rationale for the implementation of SA as a tool to promote second language (L2) writing development. The results of this study might contribute to the field by yielding empirical insights for investigating the characteristics of SA among English majors.

2 Theoretical and empirical background

2.1 Self-assessment

SA is defined as a process of formative assessment (Andrade 2019; Andrade and Du 2007), and SA practices involve learners reflecting on and evaluating the quality of their performance and their learning. Specifically, learners evaluate the extent to which their performance reflects explicitly stated objectives or criteria. SA practices also involve learners' identification of their own strengths and weaknesses in their performance, which learners revise accordingly (Andrade and Boulay 2003;

Andrade and Du 2007; Goodrich 1996; Gregory et al. 2000; Hanrahan and Isaacs 2001; Paris and Paris 2001). The focus in self-assessment is on learning and improvement as opposed to summative assessment (Andrade 2019).

Previous research has shown that SA has numerous benefits. SA increases self-awareness of learning (Babaii et al. 2016; Oscarson 1989), fosters learner autonomy (Dann 2002; Oscarson 1989), promotes self-regulated learning (Butler 2016, 2018) and motivation (Birjandi and Tamjid 2012), and reduces anxiety (Bachman and Palmer 1996). In addition, a positive association has been found between SA and learner confidence and performance (Butler and Lee 2010; De Saint Léger 2009; Little 2009). SA has also been demonstrated to bridge the gap between learner perception and actual performance (Andrade and Valtcheva 2009) and reduce the disagreement between student and teacher assessment (Babaii et al. 2016; Chen 2008). Furthermore, SA has been found to expand the range of assessment; specifically, learners can gain more profound insight into their own learning as compared to an outsider (Oscarson 1989). SA also promotes a learner-centred curriculum (Little 2009). Finally, Kato (2009) found that students considered SA activities more helpful than goal-setting activities.

However, two of the biggest concerns about SA are its validity and reliability (Ashton 2014; Patri 2002). According to Butler (2018), these concerns can be addressed by investigating the relationships between SA and objective measures of language performance. Several empirical studies have investigated the relationship between SA and language performance measurements and found positive associations between them (e.g., Ashton 2014). In Li and Zhang's (2020) meta-analysis of SA and language performance, the overall correlation between SA and language performance was moderate ($r = 0.466$, $p < 0.01$), while in an earlier meta-analysis, Ross (1998) found that the correlations ranged from $r = 0.52$ to $r = 0.65$ across the four language skills. Li and Zhang's (2020) meta-analysis also revealed that listening had the strongest correlation ($r = 0.486$), followed by reading ($r = 0.451$) and speaking ($r = 0.442$). Writing skills showed the weakest correlation ($r = 0.381$), and this is in line with Ross's (1998) results. Li and Zhang (2020) attributed the relatively weak correlation between SA and writing abilities to the features of the criteria used. While the criteria employed for listening, reading, and speaking were predominantly adopted from well-established language proficiency scales (e.g., CEFR), the writing criteria were presented using vague dimensional descriptors (e.g., topic, content, and grammar). The broad and vague writing criteria may have led to greater confusion among learners on how to interpret the criteria, which might have resulted in a large variation in SA outcomes.

2.2 Self-assessment of writing abilities

There is a positive relationship between SA and teacher assessment (TA) of writing skills (Birjandi and Tamjid 2012; Liu and Brantmeier 2019; Matsuno 2009; Saito and Fujita 2004; Summers et al. 2019; Weigle 2010; Zheng et al. 2012) and between engagement in SA and L2 writing development (Wind 2021). However, the strength of the relationships range between weak to moderate. For example, Saito and Fujita (2004) found a weak correlation between SA and TA ($r = 0.07$), while Weigle (2010) detected moderate positive correlations between SA and TA (rater 1: $r = 0.39$, rater 2: $r = 0.43$). Investigating the writing abilities and SA accuracy of 106 Chinese learners of English, Liu and Brantmeier (2019) found that young learners are also able to accurately self-assess their writing. The researchers found a significant positive relationship between SA writing and writing production ($r = 0.30$, $p < 0.01$), showing a small to medium effect size. In contrast, several studies found that SA might not be a reliable alternative for formal assessment. Matsuno (2009) found that peer-assessment can play a useful role in writing classes, whereas SA has "limited utility as a part of formal assessment" (2009: 75). Moreover, Summers et al. (2019) found weak correlations between SA and placement test results, which posits the question whether SA can be used as a placement test.

Although there have been studies investigating the relationship between SA and writing, most investigations have adopted a cross-sectional research design. Therefore, little is known about the extent to which the accuracy of SA changes over time. To the best of our knowledge, there have been two studies (Birjandi and Tamjid 2012; Zheng et al. 2012) which investigated the development of SA accuracy in the EFL context. Birjandi and Tamjid's (2012) study explored the role of SA and peer-assessment (PA) in promoting language learners' writing performance. A total of 157 English as a foreign language teacher trainees were assigned to five groups (four experimental and one control group). The participants in Group 1 used journal writing as an SA technique, while Group 2 self-assessed their performance. The participants employed PA in Group 3, whereas the participants employed both SA and PA in Group 4. In addition, TA was employed in all experimental groups with the exception of Group 4. In the control group (Group 5), there only TA was employed. The participants took a teacher-designed writing test at the beginning and another at the end of the investigation. The greatest improvements were observed in Group 2 and Group 3. Unfortunately, Birjandi and Tamjid (2012) did not give a detailed account of the writing tests used in their study. The authors stated that the participants were required to write a composition on "familiar topics" (Birjandi and Tamjid 2012: 520). However, it is not clear how many writing prompts were used and whether the same writing prompts were used at the beginning and at the end of the study.

In Zheng et al.'s (2012) study of students' SA in College English writing tests, 189 freshmen and sophomore students were instructed to assess their own writing work over an eight-week period. It was found that students could self-assess their writing quite well. The researchers highlighted that the SA of writing developed due to the instructions of the scoring rubric. After receiving rater training, the participants have shown significant ($p < 0.05$) improvement in their SA accuracy in writing. For example, the correlation increased from $r = 0.39$ to $r = 0.55$ in writing task 1, and from $r = 0.46$ to $r = 0.69$ in writing task 2. The changes were statistically significant at the $p < 0.01$ level. Therefore, their study focussing on increasing SA accuracy through training provides a solid baseline for further research. However, the order of the writing tasks in Zheng et al.'s (2012) study was not equalised. Therefore, improvements in writing might be attributed to differences in difficulty between the three writing prompts used in their study.

In conclusion, positive correlations were found between student and teacher assessment in most recent studies (Birjandi and Tamjid 2012; Liu and Brantmeier 2019; Saito and Fujita 2004; Weigle 2010; Zheng et al. 2012). However, the pitfall is that the researchers could not gain insight into temporal changes in the development of SA practices owing to the cross-sectional research designs used. In addition, studies on the SA of writing abilities used teachers' scores or human raters only and did not consider more objective measures such as linguistic complexity indices calculated by computational tools. In the next section, we will discuss the most recent findings in the field of L2 writing development.

2.3 Linguistic complexity in second language writing development

L2 writing development has generally been investigated by measuring the constructs of complexity, accuracy, and fluency. Among the three constructs, complexity is the focus of this study.

Linguistic complexity generally entails lexical and syntactic complexity. Both lexical and syntactic complexity are multidimensional constructs (Jarvis 2013; Norris and Ortega 2009). The results concerning the longitudinal developments of lexical and syntactic complexity in L2 writing are mixed. For example, Storch (2009) focused on changes in the academic writing of university students over one semester and found that participants' writing improved in structure and development of ideas but failed to improve linguistic complexity. Likewise, Knoch et al. (2015) found that clause length increased while subordination decreased over the three-year period in 32 undergraduate's writing. However, the changes were not statistically significant. Knoch et al. (2015) also found

that word length decreased while lexical sophistication increased from the pre-instructional phase (Time 1) to the post-instructional phase (Time 2); nevertheless, these changes also lacked reaching statistical significance.

In contrast, Mazgutova and Kormos (2015) found statistically significant increases in lexical variability, lexical sophistication, and cohesion over one month in the writing of an intermediate-level group, while the upper-intermediate group's writing showed significant differences only in lexical sophistication. Statistically significant increases were also found in syntactic complexity in the intermediate group's writing, whereas they only found significant differences in one syntactic complexity index in writing of the upper-intermediate group.

Although there has been an inconsistency in the definition and the operationalisation of linguistic complexity, as well as a huge variation in the duration of investigations in studies on L2 writing development, the general trend is that there are more changes at intermediate and upper-intermediate levels of proficiency (Mazgutova and Kormos 2015) than at higher levels of proficiency (Knoch et al. 2015; Storch 2009), and students appear to rely more on phrasal complexity than on subordination at advanced levels of proficiency (Halliday and Matthiessen 1999).

2.4 Research questions

Although several studies have investigated the SA of writing skills, there are few studies examining (1) whether students' SA instruction improves SA accuracy, (2) whether students' L2 writing develops as measured by TA scores and (3) whether students' L2 writing develops as measured by linguistic complexity indices. Therefore, we designed our study based on these three aims. To address the above-mentioned research niche, the present study attempts to answer the following research questions (RQs).

RQ 1: How does the relationship between self-assessment and teacher assessment total and sub-scores change over a semester-long advanced writing programme?

RQ 2: How does L2 writing change over a semester-long advanced writing programme as measured by the self-assessment and teacher assessment total scores?

RQ 3: How does L2 writing change over a semester-long advanced writing programme as measured by linguistic complexity indices?

3 Methods

3.1 Research design

This study employed an experimental design, with the experimental group receiving instruction on SA as opposed to the control group which did not receive such instruction. From the constructs of complexity, accuracy, and fluency, in our study we focused on complexity alone for two main reasons. First, accuracy was not considered in our study as the students composed their second essay electronically at Time 2. Consequently, although students were directly asked not to use any external help, we cannot rule out the possibility that they used spell-check programmes or autocorrect functions. Second, fluency, usually measured by the total number of words produced in a specific time limit, was neglected since the word count was determined by the task.

3.2 Research context

In this university, English majors are required to pass two academic skills (Academic skills 1 and 2) courses focusing on paraphrasing, summarising, and synthesising skills (Tankó 2019). These two academic skills courses are completed in the first two terms of the Bachelor of Arts (BA) in the English programme. At the end of the academic skills 1, students are required to write a guided summary, while at the end of the Academic skills 2, students are asked to write a synthesis. After completing the compulsory Academic skills courses, undergraduates are required to take the Advanced writing (AW) course aimed at improving their academic writing skills. However, in some cases there might be a year-long pause between the writing of the BA or the unified teacher training programme thesis and the completion of the AW courses. Consequently, being able to self-assess the quality of their writing might be a crucial skill for university students in Hungary.

3.3 The advanced writing course

The present research was conducted at a university in Budapest, Hungary, in AW courses during the spring term in 2020. The data for our study were collected from two AW courses taught by the authors. This, however, was not seen as an ethical issue because participation in this study was voluntary. The AW courses are usually held by different instructors, so there might be slight differences in the content, but the primary aim is to enhance students' academic writing skills mainly

by practising argumentative essay writing. The course focuses on task-based approaches involving academic reading, academic writing, critical thinking, participation in academic discussions, debating skills, receiving feedback, peer-review, self-assessment, and oral presentations. After completing weekly assignments, the students received written feedback from the instructors on how to improve their academic writing skills. The following criteria were highlighted by the instructors: forming an effective thesis statement, cohesion and coherence, paraphrasing, APA formatting of references, grammatical range and accuracy, vocabulary, quality of argumentation, style, punctuation, and paragraphing. The AW course was an ideal setting for this research for two main reasons. First, it was an intensive course focusing on writing development through detailed feedback from the instructors. Second, the participants had some prior knowledge about essay writing since they had taken the Academic skills 1 and 2 courses as prerequisites for the AW course.

3.4 Participants

The participants in this study were 33 students who enrolled in two AW courses. All of the students in the AW programme agreed to take part in our study. The students were selected by convenience and criterion sampling (Dörnyei 2007). The students were assigned to the two AW courses (AW Course 1 and AW Course 2) based on their registration in the university's system. AW Course 1 was instructed by the researcher who had no experience with SA, while the researcher who taught AW Course 2 had more experience with SA. Students in AW Course 1 (control group) did not receive SA instruction, whereas students in AW Course 2 (experimental group) received regular SA instruction.

The students were English majors around 20–25 years of age. The L1 of the participants was predominantly Hungarian ($n = 29$). However, there were also four international students (Chinese, Romanian, and Spanish). Students are eligible to register on the AW course upon successful completion of the Academic skills 1 and 2 courses. As the participants are at least third year English majors, the assumed level of their overall language proficiency was around the IELTS score of 7 (i.e., C1 based on the CEFR; Council of Europe 2001). The reason for this assumption is that first-year students at this university have to pass a Language Proficiency Exam (LPE) assessing their command of English at B2, B2+ and C1 levels as defined by the CEFR standards. The requirements and task types are based on the contents of a language practice book written by Vince and Sunderland (2003), and the exam was developed by item writers. First-year students have to complete the LPE

Table 1: The background information of the participants.

		Total	Control group	Experimental group
N		33	17	16
Gender	Female	27	14	13
	Male	6	3	3
L1 background	Hungarian	29	15	14
	Chinese	2	2	0
	Romanian	1	0	1
	Spanish	1	0	1

successfully to continue their studies. Table 1 is a summary of the participants' gender distribution and L1 background.

3.5 Instruments

The participants were asked to write two argumentative essays, one in the pre-instructional phase (Time 1) and one in the post-instructional phase (Time 2). The order of the tasks was counterbalanced; therefore, in February 2020 the first half of the participants was asked to complete Task A and the other half Task B (in both groups control and experimental). In May, after the experimental group received regular SA instruction, the participants from both groups were asked to submit the second argumentative essay. We have chosen topics related to the field of language learning as the selected participants are the most familiar with this area. The writing prompts, piloted in Wind (2018), were the following:

Task A: *A native language teacher is always better than a non-native one. To what extent do you agree?*

Task B: *The older you get, the more difficult it is to learn a foreign language. To what extent do you agree?*

Immediately after composing the 200-word-long argumentative essay, the participants were asked to self-assess their essay using a rubric based on a 5-point scale (see Table 2). The CEFR and IELTS band score equivalence of each self-assessment rubric score is also displayed in Table 2. More details regarding the instruments can be found in the Appendix. The writing rubric included the following four criteria: (1) task response, (2) coherence and cohesion, (3) vocabulary, and (4) grammatical range and accuracy. The students were asked to rate

Table 2: The assigned appropriate equivalents of SA scores with respect to IELTS band scores and CEFR levels.

Self-assessment score	IELTS band score	CEFR scale
5	9	C2
4	8	C1–C2
3	7	C1
2	6	B2
1	5	B1

The IELTS band scores are based on the descriptors of the British Council https://takeielts.britishcouncil.org/sites/default/files/ielts_task_2_writing_band_descriptors.pdf. The CEFR scale is based on the descriptors of the Council of Europe (2001).

their essays on a 5-point scale ranging from "bad" to "excellent" (1 – bad, 2 – poor, 3 – mediocre, 4 – good, 5 – excellent). At Time 1, the participants were informed about the assigned appropriate equivalents of SA scores with respect to CEFR levels. Time 1 was the pilot phase; the only issue detected by one participant concerned a spelling error in the instructions of one of the tasks. Other than that, no misunderstanding occurred in the completion of the tasks.

3.6 Data collection procedures

Data collection took place twice during the term, at the beginning and at the end of the second semester of the academic year 2019/2020. The course was planned to include 90 min of instruction per week, but this plan was disrupted by school lockdowns due to COVID-19. From mid-March, instruction and feedback were only provided online. Thus, students completed the first half of the research project in class before the pandemic and the second half through distance learning. This was not seen as a substantial drawback, however, mainly because outside class SA is not expected to differ greatly from in class SA. The students were given a writing prompt where they were asked to compose an argumentative essay of at least 200 words in 30 min. They were asked to rely on their own experience and knowledge and were not allowed to use dictionaries. However, a limitation to this might be that the use of dictionaries could not be controlled by the course instructors at Time 2. After the write-up, the students completed a writing rubric evaluating their own work immediately after they had finished.

3.7 Data analyses

The final mini corpus consisted of 66 essays of 16,920 words. We used web-based computational tools including Coh-Metrix 3.0, the L2 Syntactic Complexity Analyzer (L2SCA), and the Word and Phrase softwares to measure cohesion, syntactic complexity, and the percentage of genre-specific lexical items, respectively. We used these programmes because coding the texts manually would have been time-consuming. The texts were checked for spelling mistakes and non-existent words beforehand in order to ensure that the programmes would be able to identify and analyse the lexical items. To find the appropriate equivalents of the self-assessment scores, IELTS descriptors were used along with the CEFR scale (Council of Europe 2001).

3.7.1 Statistical analyses

First of all, we calculated normality tests; the results of the Kolmogorov–Smirnov (K–S) tests indicated that the data showed normal distribution ($p > 0.05$) with skewness and kurtosis being within the acceptable ±2 range. However, the dimensions of the assessment sub-scores showed non-normal distribution with the K–S statistic being significant ($p < 0.05$) and values for skewness and kurtosis outside the acceptable range. Although the data are normally distributed for the total scores, due to the relatively small sample size, the researchers opted for non-parametric tests. An additional reason for using non-parametric tests lies in the fact that previous studies in the field with similar research designs also used non-parametric tests to analyse small-scale data (e.g., Mazgutova and Kormos 2015).

Wilcoxon signed-rank test, the non-parametric equivalent of the paired samples *t* test, was applied to analyse the differences between the two groups. Cohen's delta was calculated using Excel to check the effect size or standardised mean difference (Cohen 1988) as it may be of crucial practical importance for researchers (Lakens 2013). It must be noted that since the data for SA in Time 2 was missing from three participants, the researchers' decision was to calculate with the SA

Table 3: Inter-rater reliability of the coding of the argumentative essays.

Phases	κ	IRR[a]	*p*	Strength of agreement
Time 1 – pre-instructional phase	0.681	84.84%	<0.001	Substantial/good
Time 2 – post-instructional phase	0.621	87.87%	<0.001	Substantial/good

[a]Inter-rater reliability based on the sum of the agreements and changes.

score of Time 1 in these three instances. This seemed to be the best option since the researchers aimed to avoid a type I error, that is, arriving at a false positive result, claiming that there were significant differences where there were none. Statistical analyses were conducted with SPSS version 22. To check inter-rater reliability, Cohen's kappa, which measures the strength of agreement between two raters or coders (Altman 1991), was calculated. As can be seen in Table 3, the inter-rater reliability of both phases reached a substantial level of agreement based on Landis and Koch (1977) and a good agreement according to Altman (1991).

3.7.2 Linguistic complexity

In this study, linguistic complexity was harmonised with the SA rubric used to score the essays. Therefore, the constructs of (1) cohesion, (2) lexical and (3) syntactic complexity were considered. Cohesion was measured by the all connectives (CNCAll) index, using the Coh-Metrix 3.0 (Graesser et al. 2004, 2011). Connectives (e.g., because, whereas, moreover) are important in creating cohesive connections between ideas, clauses, and connectives even give hints about the organisation of texts (Cain and Nash 2011; Crismore et al. 1993; Longo 1994; Sanders and Noordman 2000; van de Kopple 1985). In our study, it was expected that the incidence of all connectives might increase over time.

Lexical complexity is a multidimensional construct composed of at least three main sub-constructs: (1) lexical density, (2) lexical sophistication, and (3) lexical variability (Jarvis 2013). However, in this study we focused on the development of students' academic vocabulary. Therefore, the percentage of academic words was measured in the texts by the academic vocabulary list (AVL) index, computed by the Word and Phrase software (Gardner and Davies 2014).

Although syntactic complexity is a multidimensional construct (Norris and Ortega 2009), a general index, the mean length of clause (MLC), was calculated by the L2 Syntactic Complexity Analyzer (L2SCA; Ai and Lu 2013; Lu 2010, 2011; Lu and Ai 2015). Both Verspoor et al. (2017) and Wind (2021) have claimed that the MLC index is a reliable indicator of general syntactic complexity.

3.8 Ethical considerations and quality control

All 33 students in the course participated voluntarily in the present research project and were preliminarily informed that they had the right to opt out of the study at any time and that their anonymity was protected throughout the study. In order to ensure intercoder reliability and retain the objectivity of the analysed

texts as much as possible, Cohen's kappa was computed. The tasks piloted by Wind (2018) were piloted in the present study on the first occasion (Time 1) and proved to be understandable for the participants as no misunderstandings occurred.

4 Results and discussion

4.1 RQ 1: How does the relationship between self-assessment and teacher assessment total and sub-scores change over a semester-long advanced writing programme?

In order to answer RQ1, we first analysed the SA total scores, which were correlated with the corresponding TA total scores at the beginning (Time 1) and at the end of the AW courses (Time 2). Table 4 shows that there were positive associations between the SA and the TA total scores in both groups (control and experimental) at Time 1 and Time 2. In addition, the correlation coefficient between the SA and TA total scores was statistically significant at Time 2 in the experimental group, indicating a moderate positive relationship based on Muijs (2004) ($r = 0.502$, $p < 0.05$). Overall, SA accuracy improved in both groups over the semester-long AW programme; nevertheless; it must be noted that the improvement is statistically significant ($p < 0.05$) only in the experimental group,

Table 4: Correlations between self-assessment and teacher assessment sub-scores and total scores.

	Time 1 *r*	Time 2 *r*
Control group		
Task response	0.717[b]	0.336
Coherence and cohesion	0.314	0.373
Vocabulary	0.600[a]	0.547[a]
Grammatical range and accuracy	0.263	–0.047
Total score	0.402	0.416
Experimental group		
Task response	–0.239	0.581[a]
Coherence and cohesion	0.328	0.397
Vocabulary	0.235	0.107
Grammatical range and accuracy	0.259	–0.089
Total score	0.176	0.502[a]

[a] $p < 0.05$, [b] $p < 0.01$.

which indicates that students receiving SA instruction showed considerable improvement in their SA accuracy.

Following the analysis of the total SA scores, a correlational analysis was computed between SA and TA sub-scores at Time 1 and Time 2. Table 4 shows that there were predominantly positive relationships between the SA and TA sub-scores in both groups. There were weak negative correlations between the SA and TA scores on grammatical range and accuracy at Time 2 in both groups, and there was a weak negative association between SA and TA scores on task response at Time 1 in the experimental group. The SA-TA correlations were statistically significant ($p < 0.05$) for task response and vocabulary at Time 1 and for vocabulary at Time 2 in the control group, while the SA-TA correlation was statistically significant for task response at Time 2 in the experimental group.

The positive correlations found in our study between SA and writing performance are in line with the results of previous studies (Liu and Brantmeier 2019; Matsuno 2009; Saito and Fujita 2004; Summers et al. 2019; Weigle 2010). Nevertheless, as compared to the correlation coefficient between SA and writing skills ($r = 0.525$) reported in Ross's (1998) meta-analysis, correlations in this research endeavour are found to be weaker. In contrast, the SA-TA correlation coefficients (total scores) in the control group at Time 1 and Time 2 and in the experimental group at Time 2 in our study were stronger than the correlation coefficient between SA and writing ($r = 0.381$) reported in Li and Zhang's (2020) meta-analysis. According to Boud and Falchikov (1989), familiarity with SA might have an effect on the correlation between SA and language abilities. In our study, the relatively weak correlation coefficients might be attributed to the participants' unfamiliarity with SA practices. Thus, it can be concluded from both correlational analyses of the total and sub-scores of SA and TA that students receiving SA instruction tended to improve their SA accuracy, and this result is statistically significant.

4.2 RQ 2: How does L2 writing change over a semester-long advanced writing programme as measured by the self-assessment and teacher assessment total scores?

To answer RQ2, Wilcoxon signed-rank test was calculated to compare SA scores at Time 1 and Time 2 and the TA scores at Time 1 and Time 2. The descriptive statistics for the SA and TA total scores are displayed in Table 5. Both the SA and the TA scores increased from Time 1 to Time 2 in both groups, indicating an improvement in SA accuracy.

Table 5: Descriptive statistics for SA and TA total scores with the results of the Wilcoxon signed-rank tests for the changes in SA and TA total scores.

	Time 1		Time 2		*Z*	Cohen's *d*
	M	**SD**	***M***	**SD**		
Control group						
Total SA	3.66	1.06	3.99	0.64	−1.922	
Total TA	3.80	0.46	4.10	0.19	−2.306[a]	0.88
Experimental group						
Total SA	3.56	0.67	3.80	0.58	−1.177	
Total TA	4.02	0.31	4.18	0.18	−1.778	

[a] $p < 0.05$.

However, the results of the Wilcoxon signed-rank tests, displayed in Table 5, show statistically significant differences only for the TA scores ($Z = -2.306$, $p = 0.021$) in the control group with a large effect size ($d = 0.88$). Thus, the change from Time 1 to Time 2 in TA total scores points not only to the statistical but the practical significance (Kirk 1996) of this result. This means that the result, besides not being due to chance, may also have notable importance for writing practices.

The results of our study are also consistent with the findings of studies which investigated the development of SA accuracy in the EFL context (Birjandi and Tamjid 2012; Chen 2008; Zheng et al. 2012). Birjandi and Tamjid (2012) found that SA accuracy improved over a semester, while Chen (2008) detected development in SA accuracy over 12 weeks; Zheng et al. (2012) reported improvements in SA accuracy over an eight-week period. However, Chen (2008) focused on oral performance with two weeks of training and 10 weeks of SA and TA.

4.3 RQ 3: How does L2 writing change over a semester-long advanced writing programme as measured by the linguistic complexity indices?

To answer RQ3, Wilcoxon signed-rank tests were performed to compare linguistic complexity indices at Time 1 and Time 2. The descriptive statistics of the linguistic complexity indices are displayed in Table 6. Interestingly, the cohesion and lexical complexity indices increased in the students' essays in the control group but decreased in the students' essays in the experimental group. These results suggest that students' essays in the control group tended to become more cohesive and contain more academic words. Table 6 also demonstrates that the syntactic complexity index decreased in both groups. This result indicates that the students

Table 6: Descriptive statistics for the linguistic complexity indices.

	Control group				Experimental group			
	Time 1		Time 2		Time 1		Time 2	
	M	SD	*M*	SD	*M*	SD	*M*	SD
Cohesion								
CNCAll	98.59	18.03	102.25	14.33	93.87	18.63	93.80	19.89
Lexical complexity								
AVL	2.59	1.23	3.18	1.29	3.13	1.41	2.38	1.26
Syntactic complexity								
MLC	9.93	1.37	9.82	0.94	10.05	2.73	9.64	1.05

tended to shorten their clauses in their essays. However, the results of the Wilcoxon signed-rank tests did not show statistically significant differences for the linguistic complexity indices.

Limited changes in lexical and syntactic complexity are not infrequent in the literature on L2 writing development. For example, Knoch et al. (2015) also found no statistically significant changes in lexical and syntactic complexity measures except for fluency over a three-year degree study at a university in Australia. Likewise, Storch (2009) found no statistically significant changes in complexity and accuracy in a semester-long study at a university. The limited improvements in our study can be explained by two possible reasons. First, the duration of the investigation was relatively short compared to Knoch et al.'s (2015) three-year-long study. Second, the proficiency level of the participants was relatively high (around B2, B2+, C1 CEFR level). It can be presumed that at higher levels of language proficiency, EFL learners make improvements in fewer areas of linguistic complexity. For example, Mazgutova and Kormos (2015) found that the lower-proficiency (intermediate) group in their study improved in more areas of linguistic complexity than the higher-proficiency (upper-intermediate) group. Another possible reason for the relative stagnation of L2 writing development can be attributed to the limited functioning of self-regulatory processes, closely linked to SA. For example, Wind and Harding (2020) found that the limited use of self-regulatory processes contributed to the stagnation of the development of linguistic complexity in L2 writing.

5 Conclusions and pedagogical implications

The present results have important implications for teaching writing courses at universities as well as for language centres dedicated to improving students' writing skills.

First, our study shows that SA instruction leads to improvement in SA accuracy over a semester-long AW programme. Our results are in line with Chen's (2008) conclusions that regular feedback and practice results in improvement in learners' ability to assess their own writing. Additionally, based on the TA total scores, the students in the control group significantly developed their writing skills over a semester-long period. However, this improvement was not clearly evidenced by the changes in the linguistic complexity indices, since none of the complexity indices showed significant increases over time. Second, self-perceived weaknesses in writing (e.g., the inability to produce clear, smoothly flowing, complex essays in terms of language as well as content) can inform instructors so that they can adjust their writing instruction accordingly. Such washback effects might facilitate the promotion of learner-centred pedagogy which is particularly needed in Hungarian universities and during the COVID-19 pandemic. Teacher-centred instructions and teaching towards examinations might hinder learner autonomy and prevent students from independently setting goals and making decisions for their learning or implementing any means for reducing possible weaknesses. Along with Liu and Brantmeier (2019), we can conclude that employing SA might promote learner autonomy and university students' self-regulation.

Our study has some limitations, which should be followed up by further research. First, the number of participants (N = 33) was relatively low compared to other studies on the SA of writing skills (Birjandi and Tamjid 2012; Liu and Brantmeier 2019; Saito and Fujita 2004; Weigle 2010; Zheng et al. 2012). However, in Mazgutova and Kormos's (2015) study, the Wilcoxon signed-rank test was performed on a lower number of samples (n = 12) than in our study (n = 16). Due to the relatively low number of participants, individual differences might have encompassed some features that might have emerged in a study with a larger sample size. Consequently, future studies should replicate this research with a bigger sample.

Second, despite the fact that our findings tended to show positive correlations between SA and writing, these correlations may not entirely capture SA accuracy (Ashton 2014). The positive correlations only implied a possible trend that university students could accurately self-assess their writing performance. Additional studies on whether university students at different levels (and not only English majors) over- or under-estimate their writing skills are therefore necessary before any generalisations can be made.

Third, along with Liu and Brantmeier (2019), by only looking at the positive correlations detected in our study, we cannot verbalise how university students respond to SA items; therefore, this is yet to be examined to provide recommendations for important stakeholders (e.g., language teachers). Subsequently, further research would be indispensable for exploring the full process of SA and for understanding what leads to more accurate SA (Liu and Brantmeier 2019). Accordingly, Butler (2018) stressed that SA has a socially complex and cognitively demanding nature.

One possible future direction of research is to investigate the moderating effects of a number of variables that might play important roles in SA such as the type of criteria used in SA, the presence and form of SA criteria, SA training, the types of SA measurements, their reliability, and the number of items the SA measurement includes (Li and Zhang 2020). Furthermore, it would be worthwhile to investigate possible developments in writing and self-assessment with the same participants over at least two consecutive semesters to allow more time for improvement.

Research funding: This study was funded by the Scientific Foundations of Education Research Program of the Hungarian Academy of Sciences.

Appendix

Writing task A

Name:
Date:
You should spend about 30 min on this task. Write about the following topic:

A native language teacher is always better than a non-native one. To what extent do you agree?

Give reasons for your answer and include any relevant examples from your own knowledge or experience. Write at least 200 words.

Self-assessment

After completing the writing task, please rate your essay based on the following criteria *(5 = excellent, 4 = good, 3 = mediocre, 2 = poor, 1 = bad)*.

Task response	1	2	3	4	5
Coherence and cohesion	1	2	3	4	5
Vocabulary	1	2	3	4	5
Grammatical range and accuracy	1	2	3	4	5

Writing task B

Name:
Date:
You should spend about 30 min on this task. Write about the following topic:

The older you get, the more difficult it is to learn a foreign language. To what extent do you agree?

Give reasons for your answer and include any relevant examples from your own knowledge or experience. Write at least 200 words.

Self-assessment

After completing the writing task, please rate your essay based on the following criteria *(5 = excellent, 4 = good, 3 = mediocre, 2 = poor, 1 = bad)*.

Task response	1	2	3	4	5
Coherence and cohesion	1	2	3	4	5
Vocabulary	1	2	3	4	5
Grammatical range and accuracy	1	2	3	4	5

References

Ai, Haiyang & Xiaofei Lu. 2013. A corpus-based comparison of syntactic complexity in NNS and NS university students' writing. In Ana Díaz-Negrillo, Nicolas Ballier & Paul Thompson (eds.), *Automatic treatment and analysis of learner corpus data*, 249–264. Amsterdam & Philadelphia: John Benjamins Publishing Company.

Altman, Douglas G. 1991. *Practical statistics for medical research*. London: Chapman & Hall.

Andrade, Heidi L. 2019. A critical review of research on student self-assessment. *Frontiers in Education* 4(87). 1–13.

Andrade, Heidi G. & Beth A. Boulay. 2003. Role of rubric-referenced self-assessment in learning to write. *The Journal of Educational Research* 97(1). 21–34.

Andrade, Heidi & Ying Du. 2007. Student responses to criteria-referenced self-assessment. *Assessment & Evaluation in Higher Education* 32(2). 159–181.

Andrade, Heidi & Anna Valtcheva. 2009. Promoting learning and achievement through self-assessment. *Theory into Practice* 48(1). 12–19.
Ashton, Karen. 2014. Using self-assessment to compare learners' reading proficiency in a multilingual assessment framework. *System* 42(1). 105–119.
Babaii, Esmat, Shahin Taghaddomi & Roya Pashmforoosh. 2016. Speaking self-assessment: Mismatches between learners' and teachers' criteria. *Language Testing* 33(3). 411–437.
Bachman, Lyle F. & Adrian S. Palmer. 1996. *Language testing in practice*. Oxford: Oxford University Press.
Birjandi, Parviz & Nasrin Hadidi Tamjid. 2012. The role of self-, peer and teacher assessment in promoting Iranian EFL learners' writing performance. *Assessment & Evaluation in Higher Education* 37(5). 512–533.
Boud, David & Nancy Falchikov. 1989. Quantitative studies of student self-assessment in higher education: A critical analysis of findings. *Higher Education* 18(5). 529–549.
Butler, Yuko Goto. 2016. Assessing young learners. In Dina Tsagari (ed.), *Handbook of second language assessment*, 359–375. Berlin & New York: Mouton de Gruyter.
Butler, Yuko Goto. 2018. The role of context in young learners' processes for responding to self-assessment items. *The Modern Language Journal* 102(1). 242–261.
Butler, Yuko Goto & Jiyoon Lee. 2006. On-task versus off-task self-assessments among Korean elementary school students studying English. *The Modern Language Journal* 90(4). 506–518.
Butler, Yuko Goto & Jiyoon Lee. 2010. The effects of self-assessment among young learners of English. *Language Testing* 27(1). 5–31.
Cain, Kate & Hannah M. Nash. 2011. The influence of connectives on young readers' processing and comprehension of text. *Journal of Educational Psychology* 103(2). 429–441.
Chapelle, Carol A. & Geoff Brindley. 2010. Assessment. In Norbert Schmitt (ed.), *An introduction to applied linguistics*, 247–267. London: Hodder Education.
Chen, Yuh-Mei. 2008. Learning to self-assess oral performance in English: A longitudinal case study. *Language Teaching Research* 12(2). 235–262.
Cohen, Jacob. 1988. *Statistical power analysis for the behavioral sciences*, 2nd edn. New York: Lawrence Erlbaum Associates Publishers.
Council of Europe. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. Strasbourg: Language Policy Unit. Available at: https://rm.coe.int/16802fc1bf.
Crismore, Avon, Raija Markkanen & Margaret S. Steffensen. 1993. Metadiscourse in persuasive writing: A study of texts written by American and Finnish university students. *Written Communication* 10(1). 39–71.
Csépes, Ildikó. 2016. Language assessment literacy in English teacher education in Hungary. In Dina Tsagari (ed.), *Classroom-based assessment in L2 contexts*, 30–53. Newcastle upon Tyne: Cambridge Scholar Publishing.
Dann, Ruth. 2002. *Promoting assessment as learning: Improving the learning process*. London: Routledge.
De Saint Léger, Diane. 2009. Self-assessment of speaking skills and participation in a foreign language class. *Foreign Language Annals* 42(1). 158–178.
Dörnyei, Zoltán. 2007. *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. Oxford: Oxford University Press.
Gardner, Dee & Mark Davies. 2014. A new academic vocabulary list. *Applied Linguistics* 35(3). 305–327.

Goodrich, Heidi Watts. 1996. *Student self-assessment: At the intersection of metacognition and authentic assessment*. Cambridge: Harvard University Dissertation.
Graesser, Arthur C., Danielle S. McNamara, Max M. Louwerse & Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers* 36(2). 193–202.
Graesser, Arthur C., Danielle S. McNamara & Jonna M. Kulikowich. 2011. Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher* 40(5). 223–234.
Gregory, Kathleen, Caren Cameron & Anne Davies. 2000. *Self-assessment and goal-setting*. Courtenay: Connection Publishing.
Halliday, Michael A. K. & Christian M. I. M. Matthiessen. 1999. *Construing experience through meaning: A language-based approach to cognition*. London: Cassell.
Hanrahan, Stephanie J. & Geoff Isaacs. 2001. Assessing self- and peer-assessment: The students' views. *Higher Educational Research and Development* 20(1). 53–70.
Hasselgreen, Angela. 2000. The assessment of the English ability of young learners in Norwegian schools: An innovative approach. *Language Testing* 17(2). 261–277.
Hubai, Katalin & Ildikó Lázár. 2018. Assessment of learning in the Hungarian education system with a special focus on language teachers' views and practices. *Working Papers in Language Pedagogy* 12. 83–93.
Jarvis, Scott. 2013. Defining and measuring lexical diversity. In Scott Jarvis & Michael Daller (eds.), *Vocabulary knowledge: Human ratings and automated measures*, 13–45. Amsterdam & Philadelphia: John Benjamins Publishing Company.
Kato, Fumie. 2009. Student preferences: Goal-setting and self-assessment activities in a tertiary education environment. *Language Teaching Research* 13(2). 177–199.
Kirk, Roger E. 1996. Practical significance: A concept whose time has come. *Educational and Psychological Measurement* 56(5). 746–759.
Knoch, Ute, Amir Rouhshad, Su Ping Oon & Neomy Storch. 2015. What happens to ESL students' writing after three years of study at an English medium university? *Journal of Second Language Writing* 28. 39–52.
Lakens, Daniël. 2013. Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology* 4(863). 1–12.
Landis, J. Richard & Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33(1). 159–174.
Li, Minzi & Xian Zhang. 2020. A meta-analysis of self-assessment and language performance in language testing and assessment. *Language Testing* 38(2). 189–218.
Little, David. 2009. Language learner autonomy and the European language portfolio: Two L2 English examples. *Language Teaching: Surveys and Studies* 42(2). 222–233.
Liu, Huan & Cindy Brantmeier. 2019. "I know English": Self-assessment of foreign language reading and writing abilities among young Chinese learners of English. *System* 80. 60–72.
Longo, Bernadette. 1994. The role of metadiscourse in persuasion. *Technical Communication* 41(2). 348–352.
Lu, Xiaofei. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics* 15(4). 474–496.
Lu, Xiaofei. 2011. A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly* 45(1). 36–62.
Lu, Xiaofei & Haiyang Ai. 2015. Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing* 29. 16–27.

Matsuno, Sumie. 2009. Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing* 26(1). 75–100.
Mazgutova, Diana & Judit Kormos. 2015. Syntactic and lexical development in an intensive English for Academic Purposes programme. *Journal of Second Language Writing* 29. 3–15.
Muijs, Daniel. 2004. *Doing quantitative research in education with SPSS*. London: Sage Publications.
Norris, John M. & Lourdes Ortega. 2009. Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics* 30(4). 555–578.
Oscarson, Mats. 1989. Self-assessment of language proficiency: Rationale and applications. *Language Testing* 6(1). 1–13.
Paris, Scott G. & Alison H. Paris. 2001. Classroom applications of research on self-regulated learning. *Educational Psychologist* 36(2). 89–101.
Patri, Mrudula. 2002. The influence of peer feedback on self- and peer-assessment of oral skills. *Language Testing* 19(2). 109–131.
Ross, Steven. 1998. Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing* 15(1). 1–20.
Saito, Hidetoshi & Tomoko Fujita. 2004. Characteristics and user acceptance of peer rating in EFL writing classrooms. *Language Teaching Research* 8(1). 31–54.
Sanders, Ted J. M. & Leo G. M. Noordman. 2000. The role of coherence relations and their linguistic markers in text processing. *Discourse Processes* 29. 37–60.
Summers, Maria M., Troy L. Cox, Benjamin L. McMurry & Dan P. Dewey. 2019. Investigating the use of the ACTFL can-do statements in a self-assessment for student placement in an Intensive English Program. *System* 80. 269–287.
Storch, Neomy. 2009. The impact of studying in a second language (L2) medium university on the development of L2 writing. *Journal of Second Language Writing* 18(2). 103–118.
Tankó, Gyula. 2019. *Paraphrasing, summarising and synthesising skills for academic writers: Theory and practice*, 2nd edn. Budapest: Eötvös University Press.
Vande Kopple, William J. 1985. Some exploratory discourse on metadiscourse. *College Composition & Communication* 36(1). 82–93.
Verspoor, Marjolijn, Wander Lowie, Hui Ping Chan & Louisa Vahtrick. 2017. Linguistic complexity in second language development: Variability and variation at advanced stages. *Recherches en didactique des langues et des cultures: Les cahiers de l'Acedle* 14(1). 1–27.
Vince, Michael & Peter Sunderland. 2003. *Advanced language practice: English grammar and vocabulary*. Oxford: Macmillan Education.
Weigle, Sara Cushing. 2010. Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing* 27(3). 335–353.
Wind, Attila M. 2018. *Second language writing development from a complex dynamic systems theory perspective: A multiple case-study of Hungarian learners of English*. Lancaster: Lancaster University PhD Thesis. https://doi.org/10.17635/lancaster/thesis/519.
Wind, Attila M. 2021. Nonlinearity and inter- and intra-individual variability in the extent of engagement in self-reflection and its role in second language writing: A multiple-case study. *System* 103. 102672.
Wind, Attila M. & Luke Harding. 2020. Attractor states in the development of linguistic complexity in second language writing and the role of self-regulation: A longitudinal case study. In Wander Lowie, Marije Michel, Audrey Rousee-Malpat, Merel Keijzer & Rasmus Steinkrauss

(eds.), *Usage-based dynamics in second language development*, 130–154. Bristol: Multilingual Matters.

Zheng, Huiqing, Jianbin Huang & Ying Chen. 2012. Effects of self-assessment training on Chinese students' performance on college English writing tests. *Polyglossia* 23. 33–42.