Nordand

VITENSKAPELIG PUBLIKASJON

Scandinavian University Press

Årgang 20, nr. 1-2025, s. 1–17
ISSN online: 2535-3381
DOI: https://doi.org/10.18261/nordand.20.1.1

Selecting Instruments for Investigating L2 Swedish Writing Data of L1 Dutch Learners Across Proficiencies

Margot Fonteyne
Doctoral Researcher, Ghent University
Margot.Fonteyne@UGent.be
https://orcid.org/0000-0002-9575-7844

Maribel Montero Perez
Assistant Professor, Ghent University
Maribel.MonteroPerez@UGent.be
https://orcid.org/0000-0002-0868-588X

Joke Daems
Assistant Professor, Ghent University
Joke.Daems@UGent.be
https://orcid.org/0000-0003-3734-5160

Lieve Macken
Associate Professor, Ghent University
Lieve.Macken@UGent.be
https://orcid.org/0000-0001-7516-7487

Abstract

When conducting second language (L2) writing experiments with participants of varying proficiency, careful selection of both the test used to estimate participants' L2 proficiency and the tasks used to elicit writing data from participants is essential. However, L2 writing studies involving learners with varying proficiency rarely discuss the criteria or rationale behind the selection of their research instruments. In this article, we describe how we collected data from a small sample of L1 Dutch learners of Swedish with a wide range of proficiencies and used these to inform our selection of a proficiency test and a set of writing tasks for a research project on the role of proficiency in L2 writing. Each learner completed two proficiency tests, four self-developed writing tasks, and a post-task interview. We used these data to examine the suitability of the two proficiency tests for estimating participants' L2 Swedish proficiency and the ability of the writing tasks to elicit data from all participants, regardless of proficiency. One proficiency test outperformed the other in terms of efficiency. The writing tasks elicited suitable writing data from all participants, although the interview data also revealed some aspects of the writing tasks that could be improved.

Keywords
L2 writing, Swedish, proficiency estimation, writing tasks

Introduction

Over the years, interest in understanding the role of L2 proficiency in L2 writing has grown (e.g., Révész et al., 2022; Vasylets & Marín, 2021). Likewise, research on L2 writing in languages other than English has increased (Sun & Lan, 2023). Experimental studies in these areas typically involve collecting writing data from participants with a wide range of proficiencies. In studies on the role of L2 proficiency in L2 writing, participants of varying proficiency are recruited to ensure that the variable of interest displays sufficient variability to address the research questions. In studies on L2 writing in languages other than English, learner populations can be so small that a large sample of participants with similar proficiencies may be difficult to recruit.

During instrument selection, L2 writing experiments that involve participants of varying proficiency commonly face two needs:

1) An L2 proficiency test is needed to assess the impact of L2 proficiency on writing or to control for proficiency during data analysis. For languages other than English, selecting such a test is often less straightforward. While general conventions may exist for these languages, there is often no clear standard for which proficiency test to use specifically for experimental studies on L2 writing.
2) A suitable set of writing tasks is essential to elicit writing data from participants of varying proficiency. Depending on the research aims, these tasks should facilitate the investigation of the final written products and/or the writing processes, including real-time behaviours (e.g., pausing and revising) and underlying cognitive activities (e.g., planning, formulating, and evaluating), across proficiencies.

These two needs also arose during instrument selection for a research project involving learners of Swedish of varying proficiency. The project aims to assess the impact of online tools (e.g., dictionaries, machine translation tools) on the writing products and processes of L1 Dutch learners of Swedish and whether the impact of these tools varies depending on learners' L2 proficiency. Hence, the project involves learners of varying proficiency. Deciding on which test to use as an L2 proficiency estimator in this project proved difficult, because no studies exist that examine the various tests available for estimating L2 Swedish proficiency. In addition, we were hesitant to reuse writing tasks from other L2 writing studies with learners of varying proficiency, because these studies rarely discuss the suitability of the selected writing tasks to elicit data across proficiencies. We therefore decided to set up a preliminary study, in which we investigate the suitability of two carefully selected L2 Swedish proficiency tests and a set of four self-developed writing tasks for use with learners of varying proficiency. This article reports on the findings of this study.

Background

Estimating L2 Swedish Proficiency for (Writing) Research Purposes

In research on second language acquisition (SLA), it is not uncommon to estimate proficiency using learners' classroom level, year of instruction, or self-assessments (Park et al., 2022). However, although convenient, such measures tend to be imprecise. Therefore, Park et al. (2022) argue that SLA researchers should measure participants' proficiency with independent tests, such as standardised placement tests or tests that yield proxy measures of L2 proficiency. Examples of the latter are cloze tests or C-tests, oral tests (e.g., elicited imitation tasks), or vocabulary tests (e.g., vocabulary levels tests or lexical decision tasks). Although they do not directly measure L2 proficiency, they do measure constructs that are strongly correlated with it (Kostromitina & Plonsky, 2022; Milton, 2013). The scores on these tests can therefore function as proxy measures of L2 proficiency. Following the recommendations of Park et al. (2022), we scanned prior studies on L2 Swedish involving participants of varying proficiency and academic literature documenting test development and design for references to such tests for L2 Swedish learners. We found two standardised placement tests and two vocabulary levels tests.

The first placement test is the online DIALANG test for Swedish[1] (Huhta et al., 2002). It targets various components of L2 proficiency through self-assessment, multiple-choice and short-answer questions. For each component, the test taker receives a numerical score and/or a Common European Framework (CEFR) level estimate (Council of Europe, 2020), but no overall score is given. Completing the test takes about two hours, but it is also possible to only complete certain sections. Several studies on L2 Swedish (Długosz, 2023; Tiselius & Sneed, 2020) have used the test to estimate participants' L2 proficiency.

The second placement test is a paper-based test that we will refer to as the Folkuniversitetet Test (FUT), because it was developed by the Swedish adult educational association *Folkuniversitetet*. Three different versions of the test exist (A, B, and C), each tailored to a different proficiency range (pre-A1 to A2+, A2+ to B2, and B2+ to C1+). Each version contains 40 gap-filling questions testing learners' grammatical and lexical competence. Version B, for example, is split up into three parts. The first part consists of 19 multiple-choice gap-filling questions evaluating mostly grammatical competence (e.g., prepositions, conjunctions, agreement, word order). The second section comprises 10 open-ended cloze sentences covering mainly vocabulary knowledge. The last part of 11 questions captures participants' verb knowledge (e.g., form, tense, and voice) in the form of gap-filling exercises. The result is a score out of 40. Its duration is unknown. Prentice and Forsberg Lundell (2021) use version C in their study with upper-intermediate to advanced learners of Swedish.

The first of the two vocabulary tests is the Swedish Levels Test (SweLT). Its development and validation have been thoroughly documented in Bokander (2016). The SweLT was created in both paper and digital formats and is divided into four sections of 20 items, intended to measure learners' receptive vocabulary knowledge at four levels (the 2,000, 3,000, 5,000, and 8,000 most common Swedish words). Each item consists of a gapped sentence and four words sampled from a frequency-ranked base vocabulary pool (Forsbom, 2006) extracted from the Stockholm-Umeå corpus (1M tokens) (Källgren et al., 2006). Test takers are supposed to select the word that fits best into the sentence. The result is a score out of 80. There is no information available on its exact duration, but the test administrators in Bokander's study perceived 25 to 30 minutes as sufficient.

The second vocabulary test is the paper-based Swedish Levels Test (SLT) constructed by Lindberg and Johansson (2018). The SLT consists of 48 items, each comprising six words and three definitions. The words are taken from a frequency list of a self-compiled corpus of normative Swedish texts (285M+ tokens) and from the Swedish Academic Vocabulary List (Ribeck et al., 2014), but the exact sampling procedure remains unclear. For each item, test takers are supposed to match three words to their respective definitions. The result is a score out of 288. Its average duration is not mentioned. Lindberg and Johansson (2018) used the SLT to measure L1 and L2 Swedish high school students' vocabulary breadth.

Eliciting Writing Data from Learners of Varying Proficiency

Designing tasks to elicit writing data from L2 learners involves several decisions, including genre, topic, text length requirements, and time constraints. Experimental L2 writing studies with participants of varying proficiency have employed a range of genres. Examples can be found of narrative (Kim et al., 2016; Lu, 2020; Vasylets & Marín, 2021), descriptive (Liao, 2022), and expository (Barkaoui, 2019; Lahuerta, 2018) writing tasks, but argumentative writing tasks appear to be most prevalent (Gánem-Gutiérrez & Gilmore, 2018; Khushik & Huhta, 2020; Révész et al., 2022; Tiryakioglu et al., 2019). The topics selected for these tasks are equally diverse, ranging from birthday presents (Tiryakioglu et al., 2019) to immigration politics (Vallejos, 2020). Regarding text length, studies requiring a certain length often ask participants to write 200 to 300 words (Barkaoui, 2019; Vallejos, 2020; Vasylets & Marín, 2021). Others set a minimum word count (Khushik & Huhta, 2020; Vasylets & Marín, 2021) or expect participants to stay within a certain range (Barrot & Agdeppa, 2021; Lahuerta, 2018). In terms of time allotment, most studies impose strict limits. A maximum of 30 minutes is the most common (Kim et al., 2016; Lu, 2020; Révész et al., 2022; Tiryakioglu et al., 2019). Other studies offer more flexibility by setting a time range (Barrot & Agdeppa, 2021) or allowing extensions (Sasaki, 2000).

How suitable these various genres, topics, text length requirements, or time constraints are for eliciting data from learners of varying proficiency is rarely discussed, although some exceptions exist. Lu (2020), for example, relied on expert ratings to determine which tasks from a set of potential writing tasks were of average difficulty and therefore appropriate for participants of varying proficiency. In addition, no word count was set to accommodate for the variation in L2 proficiency among participants. Another example is Kim et al. (2016), who used non-verbal videos as source material for their story retelling tasks, so that "all learners could view and understand the clips, regardless of […] their Korean proficiency level" (p. 160). Mock evaluations confirmed that the videos did not raise comprehensibility issues. Lastly, Sasaki (2000) and Vallejos (2020) based the time limits for their writing tasks on how much time L2 writers needed to complete these tasks in an untimed pilot study.

Research Objectives

In our literature review, we identified several tests that could be used for estimating participants' proficiency in L2 Swedish (writing) research. However, there is limited information on how these tests compare. This article aims to provide a first comparison of two of these tests – that is, the SweLT and version B of the FUT (FUT B) – regarding their suitability as L2 proficiency estimators in experimental L2 Swedish (writing) research. We focus on the SweLT and the FUT B because they have several characteristics that lead us to believe that they are particularly suited for this purpose. First, they are quick to administer, counting only 80 and 40 items, respectively. Second, their length is fixed. Third, they yield a single, fine-grained score of L2 proficiency. Fourth, they target components of L2 proficiency – that is, lexical and grammatical competence – that correlate strongly with writing ability (Kojima et al., 2022) and are therefore particularly relevant for research on L2 writing. Fifth, compared to FUT A and C, we expect FUT B (tailored to A2+ to B2 learners) to have a lower risk of floor and ceiling effects when used with participants of widely varying proficiency.

Furthermore, the literature review indicates that L2 writing studies involving participants of varying proficiency rarely discuss the suitability of the selected writing tasks to elicit data across proficiencies. In response, the second objective of this article is to develop a set of writing tasks and provide a comprehensive report on the ability of these tasks to elicit such data. We will address not only the suitability of the writing tasks to elicit product data across proficiencies, but also to elicit process data, including interactions between learners and online tools. In addition, we will report on how learners of varying proficiency perceive the writing tasks.

In short, our first aim is to evaluate the suitability of the SweLT and FUT B as assessment instruments, while our second aim is to investigate the suitability of the writing tasks as elicitation instruments.

Methodology

Participants

Eleven L2 learners of Swedish (six female, three male, two non-binary) enrolled at a Flemish university participated in the study. All participants were between 18 and 24 years old and were L1 speakers of Dutch, except for one participant with an advanced level of Dutch. Their L2 proficiency varied widely. Our sample ranged from learners who had recently completed the university's Swedish proficiency course for beginners (situated at the CEFR A1 level) to learners enrolled in graduate courses in Swedish linguistics (with C1 as a prerequisite for enrolment). We recruited by posting a call for participants in Facebook groups and asking lecturers to forward the call via mail. Participation was entirely voluntary, and participants were rewarded with money (€62.50).

Proficiency Tests

We digitised both the FUT B and the SweLT in Google Forms, using short-answer and multiple-choice question types, and translated the instructions into Dutch.[2] We did not set any time limits, as the original proficiency tests were not timed either, but we did add *I don't know* as a response option to all multiple-choice questions to discourage guessing behaviour (Zhang, 2013). Lastly, we noticed that the SweLT contained some (near) duplicate words in the response options, so we replaced these with alternatives.[3]

We also added answer keys to Google Forms so participants' scores could be calculated automatically. Answers on the SweLT were scored by awarding 1 point for each correctly selected option, with 80 being the maximum score possible. For the multiple-choice questions in the FUT B, we adopted the same approach. For the open-ended questions, however, we adopted an exact-scoring approach: A point was awarded solely when the provided answer exactly matched the solution. The maximum score possible on the FUT B was 40.

Writing Tasks

We developed a set of four writing tasks. Each writing task is designed to elicit two types of writing: Participants are first asked to describe three pictures (descriptive writing) and then explain which one appeals to them the most (expository writing). According to the CEFR (Council of Europe, 2020), A1 learners can describe familiar objects in simple language, while expressing opinions is only expected from B1. We therefore expect the writing tasks to be fairly manageable for lower-proficiency learners while also offering opportunities for higher-proficiency learners to demonstrate their skills.

We chose picture-based writing tasks for four reasons. First, pictures provide writers with content support, which should limit the impact of individual differences in topic knowledge (Polio & Friedman, 2016). Second, unlike L2 input, picture-based input does not allow learners to reuse fragments from the instructions. This ensures that the writing products accurately represent learners' skills in the L2 (Alanen et al., 2010; Sasayama et al., 2021). Third, picture-based input is comprehensible to all learners (Kim et al., 2016). Fourth, pictures help control the content of the writing products (Coyle et al., 2023), which should make them more comparable. In addition, we consider pictures to be particularly suitable for eliciting tool interactions, as learners are unlikely to be familiar with all the words needed to describe the pictures.

Depending on the writing task, participants are asked to describe pictures of three different events, holidays, travel destinations, or workshops (see Table 1). All pictures are connected to the Nordics or Nordic culture. For instance, the 'events' writing task contains a picture of an ice hockey cup, a picture of the song contest *Melodifestivalen*, and a picture of the Nobel Prize ceremony. We chose the Nordics because we assume it is a topic that generally interests learners of Swedish, which should motivate them to perform well on the writing tasks (Weigle, 2010). We also wanted the writing tasks to prompt even the higher-proficiency participants to regularly resort to tools, so the elicited data would allow us to investigate how tools affect learners' writing across all proficiencies. We therefore selected pictures whose description demands knowledge of rather low-frequency vocabulary (e.g., *sticknål* (knitting needle), *skärgård* (archipelago), *midsommarstång* (maypole), or *hockeyklubba* (hockey stick)).

Table 1. Overview of the pictures included in the writing tasks

Picture	Writing task topics			
	Workshops	Travel destinations	Holiday celebrations	Events
1	Norwegian knitting	Swedish inland	*Luciadagen* (Saint Lucy's Day)	Ice hockey cup
2	Ice sculpting	Major city in Sweden	*Midsommar* (Midsummer)	*Melodifestivalen* (song contest)
3	Baking cinnamon rolls	Swedish archipelago	*Valborg* (Walpurgis Night)	Nobel Prize award ceremony

To ensure that writing instructions were clear, we provided them in the participants' L1 (Dutch) (see Appendix for an English translation). We included time constraints, because we adhered to a fixed payment scheme and wanted to avoid unfair compensations due to substantial variation in completion time across participants. In addition, timed conditions help to increase the comparability of participants' writing. Following similar studies (see above), participants were allowed a 30-minute window to complete their text, with the option to submit earlier if they considered it finished. The instructions also told participants to aim for a half-page essay, which amounted to roughly 400 words in the Word template provided to the participants. We opted for this loose goal rather than a strict requirement (i.e., "aim for" instead of "produce"), so the less proficient would not feel too pressured to reach this text length. The instructions also emphasised that the descriptive and expository parts should be of approximately equal length, regardless of the final text length.

Our analysis of the writing tasks in this article is based on the data collected from the final seven participants in our sample of 11 learners. Although we also collected writing data from the first four participants, the set of writing tasks we used with these participants was still under development. Specifically, this set did not specify a text length and only featured two pictures per writing task.

Procedure

All participants individually attended two data collection sessions, each lasting approximately two hours. In the first session, they completed two of the four writing tasks, the SweLT, and a background questionnaire asking for their year of instruction and self-assessments of their Swedish writing, reading, speaking, and listening skills. In the second session, participants completed the remaining two writing tasks, the FUT B, and a post-task interview in Dutch about their perceptions of the writing tasks. Each session also included two 10-minute breaks. To reduce order, learning, and carry-over effects, the writing tasks were assigned to the participants in four different sequences.

Before each writing task, participants were provided time to read the instructions. They completed the writing tasks in Microsoft Word, which enabled us to collect data on their writing processes with the keystroke logging software Inputlog[4] (Leijten & Van Waes, 2013). To facilitate tool interactions, they could consult a machine translation tool (DeepL) for two of the writing tasks and an online bilingual dictionary (Van Dale) for the other two. We alternated the order in which these tools were assigned to the writing tasks. Participants could only translate between Swedish and Dutch and were not allowed to use any other tools. Word's built-in spelling and grammar checker was disabled by default, and participants did not activate it.

The procedure was approved by the ethical committee of the authors' institution. The author who collected the data did not know the participants.

Data Analysis

To investigate the suitability of the FUT B and the SweLT in estimating the L2 Swedish proficiency of our participants, we analysed participants' scores on the proficiency tests, the time they needed to complete them, and the relationship of the scores to two other proficiency measures: their self-assessments and the year of L2 Swedish instruction in which they were enrolled. Participants' self-assessments of their Swedish writing, reading, speaking, and listening skills were aggregated into composite scores. As each skill was rated on a five-point scale, the highest possible composite score was 20. Participants' composite self-assessment scores and their proficiency test scores were converted into percentages.

To investigate the suitability of the writing tasks for eliciting data from participants of varying proficiency in experimental L2 writing research, we selected five measures: text length, process duration, percentage of expository writing, tool consultation frequency, and percentage of tool use.

We used text length (in words) to investigate whether the writing tasks elicit sufficiently long writing products across proficiencies. Previous research recommends a minimum of 100 words for computing measures of lexical (Zenker & Kyle, 2021) and syntactic (Hwang & Polio, 2023) complexity, which are two constructs widely used for analysing L2 production (Bulté & Housen, 2012). Therefore, it is important that the writing tasks we designed allow learners to produce at least 100 words within the given time limit. However, reaching this target may be more challenging for less proficient learners, who tend to write more slowly (Barkaoui, 2019; Révész et al., 2022; Spelman Miller et al., 2008).

We used process duration (in minutes) to determine whether the writing tasks elicit full writing processes across proficiencies. Although there are currently no established thresholds for the amount of process data (in keystrokes or minutes) needed to compute common writing process measures, L2 writers are known to display different behaviours and engage in different subprocesses at various stages of writing (e.g., more revising towards the end) (Gánem-Gutiérrez & Gilmore, 2018; Révész et al., 2022). Thus, the writing tasks should allow learners to complete a full writing process within the given time limit. Again, this may be more challenging for less proficient learners, who tend to write more slowly.

We used the percentage of expository writing to examine whether the writing tasks elicit similar amounts of descriptive and expository writing across proficiencies. Since genre has been shown to affect L2 writing products and processes (Lu, 2020; Yoon & Polio, 2017), it is important for the writing tasks to elicit a balanced mix of descriptive and expository writing. However, as expressing opinions is typically only expected from more proficient learners (Council of Europe, 2020), producing expository writing may be more challenging for less proficient learners.

We used tool consultation frequency (number of times participants opened up a tool per minute) and percentage of tool use (time spent in tools divided by process duration) to investigate whether the writing tasks can elicit sufficient tool interactions across proficiencies. Previous studies have shown that more proficient learners tend to rely less on online tools while writing (Gánem-Gutiérrez & Gilmore, 2018; Xu & Ding, 2014). Therefore, writing tasks that are too simple may not elicit sufficient tool interactions from higher-proficiency learners.

For the text length, process duration, tool consultation frequency, and percentage of tool use, we relied on the summary statistics Inputlog provided for each keystroke log. For the percentage of expository writing, we manually identified the expository paragraphs in the texts, counted the number of words, and divided these counts by the total number of words in the text. For each of the five measures, we calculated the mean per participant across the four writing tasks and then computed descriptive statistics for those means. Additionally, to analyse participants' perceptions regarding the suitability of the writing tasks, we first verbatim transcribed the post-task interviews. Then, we identified common themes in participants' answers to the interview question: "Did you find the writing tasks suitable for you, given your L2 Swedish proficiency, or too easy or too difficult? Why?"

Results

Proficiency Tests

Table 2 provides descriptives for participants' proficiency test scores and completion times. Participants' mean SweLT score is comparable to the mean reported by Bokander (2016) (71%, *N* = 290) and noticeably higher than the mean FUT B score. Note, however, that the confidence intervals overlap. The SweLT scores also appear to be less dispersed than the FUT B scores, given their smaller standard deviation and narrower range. While the lowest FUT B score is closer to the minimum than the lowest SweLT score, the highest SweLT score is slightly closer to the maximum than the highest FUT B score. Regarding completion times, participants needed more time to complete the SweLT than the FUT B.

Table 2. Descriptive statistics for the SweLT and FUT B scores and completion times (*N* = 11)

Statistic	Score (%)		Completion time (in minutes)	
	SweLT	FUT B	SweLT	FUT B
M	69	52	18.45	11.18
SD	16	21	5.84	4.21
95% CI	[58, 79]	[38, 66]	[14.53, 22.38]	[8.35, 14.01]
Range	41–89	20–80	13.00–31.00	5.00–17.00

Figure 1 visualises participants' scores on the two proficiency tests, their composite self-assessment scores and year of instruction. The graph suggests a positive relationship between the FUT B and SweLT scores. Although the limited sample size warrants caution, Spearman's rank correlation coefficient supports this observation, with *r*(9) = .72. However, the graph does not show a clear connection between participants' proficiency test scores and their self-assessments. Spearman's rank correlation between the FUT B score and self-assessments is negligible, *r*(9) = .03, but moderate between the SweLT score and self-assessments, $r(9) = .59$. Lastly, participants' scores do seem to increase with higher years of instruction. However, this relationship is less evident than the one between the FUT B and SweLT scores. Spearman's rank correlations suggest moderate relationships: $r(9) = .52$ for FUT B score and year of instruction, and $r(9) = .48$ for SweLT score and year of instruction.

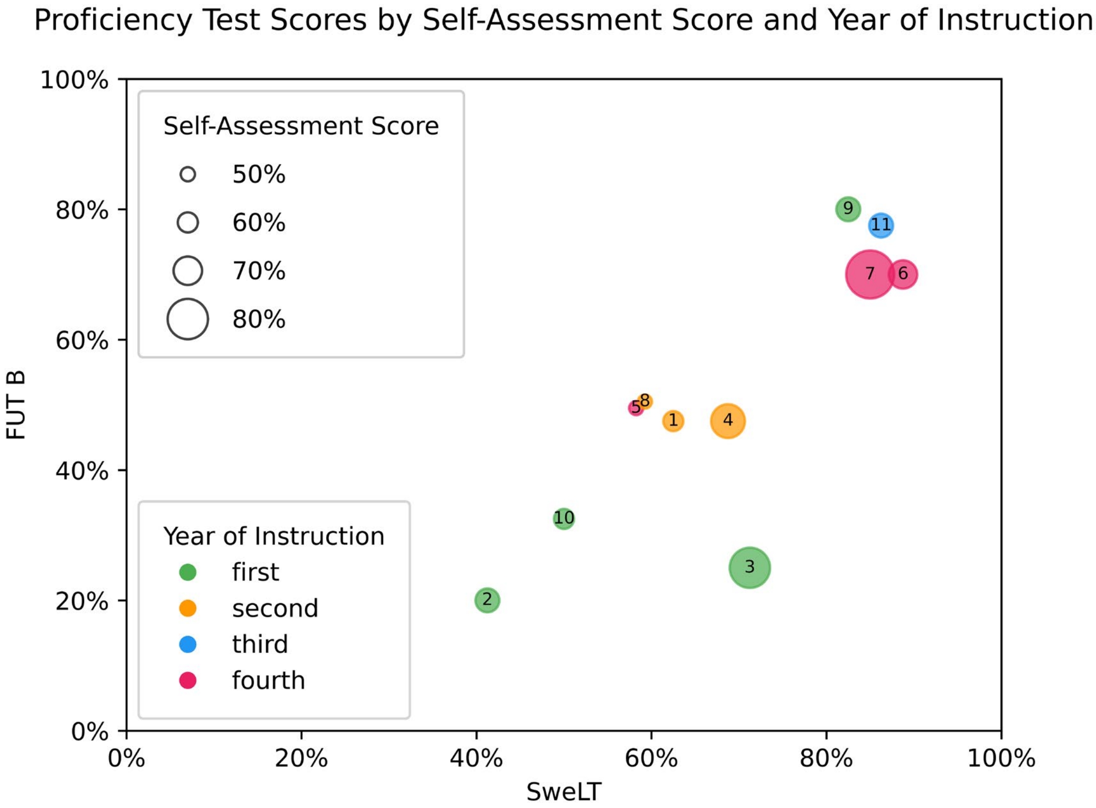

Figure 1. Scatter plot showing participants' FUT B and SweLT scores, with sizes representing composite self-assessment scores and colours representing year of instruction. Each data point is labelled with the participant's ID. To improve visibility, the identical scores for Participant 5 and 8 have been adjusted slightly.

As can also be seen in Figure 1, Participants 3, 5, 9, and 11 notably deviate from the trend of higher scores with more years of instruction. Participant 3, a first-year student, scored low on the FUT B (25%, comparable to Participants 2 and 10, two other first-year students), but high on the SweLT (71%, slightly above Participants 1, 4, and 8, all second-year students). This anomaly may be due to the participant's long completion time for the SweLT (31 minutes, 1 standard deviation above the second-highest time and 2.5 standard deviations above the mean). In contrast, they completed the FUT B in 16 minutes, which, while also on the longer end, is only about 1 standard deviation above the mean. Participant 5 scored notably lower than the other two fourth-year students, Participants 6 and 7. The self-assessments suggest that Participant 5 may be a weaker fourth-year student. Participant 5's composite self-assessment score is only 50%, compared to 70% and 85% for Participants 6 and 7, respectively. Participant 9, a first-year student, scored exceptionally high on both the FUT B and the SweLT. This may be because this participant spent a year in Sweden before starting formal L2 Swedish instruction in Flanders. Lastly, Participant 11, the sole third-year student, obtained scores very similar to fourth-year students Participants 6 and 7. This is likely because Participant 11 participated in the study in the second semester of their third year, while Participants 6 and 7 participated in the first semester of their fourth year, resulting in only a half-year difference.

Writing Tasks

The statistics for text length, process duration, and percentage of expository writing are presented in Table 3. As the statistics indicate that participants' process duration frequently approached the 30-minute time limit, we visualized the process durations of each participant in Figure 2. The graph shows that participants rarely needed the full 30 minutes to complete a writing task, except for Participant 10, whose average duration was 29.98 minutes. However, despite their lower proficiency, they produced an average of 274 words per task, with a balanced mix of descriptive and expository writing (54% expository on average). Moreover, in the post-task interview, the participant did not report any issues with completing the writing tasks. This suggests that they also managed to complete full writing processes within the given time limit.

Table 3. Descriptive statistics for text length, duration, and percentage of expository writing (*N* = 7)

Statistic	Text length (in words)	Duration (in minutes)	Percentage of expository writing
M	278.64	26.94	48.72
SD	67.16	2.20	4.35
95% CI	[216.53, 340.76]	[24.92, 28.98]	[44.70, 52.75]
Range	162.00–359.00	23.09–29.83	40.48–54.00

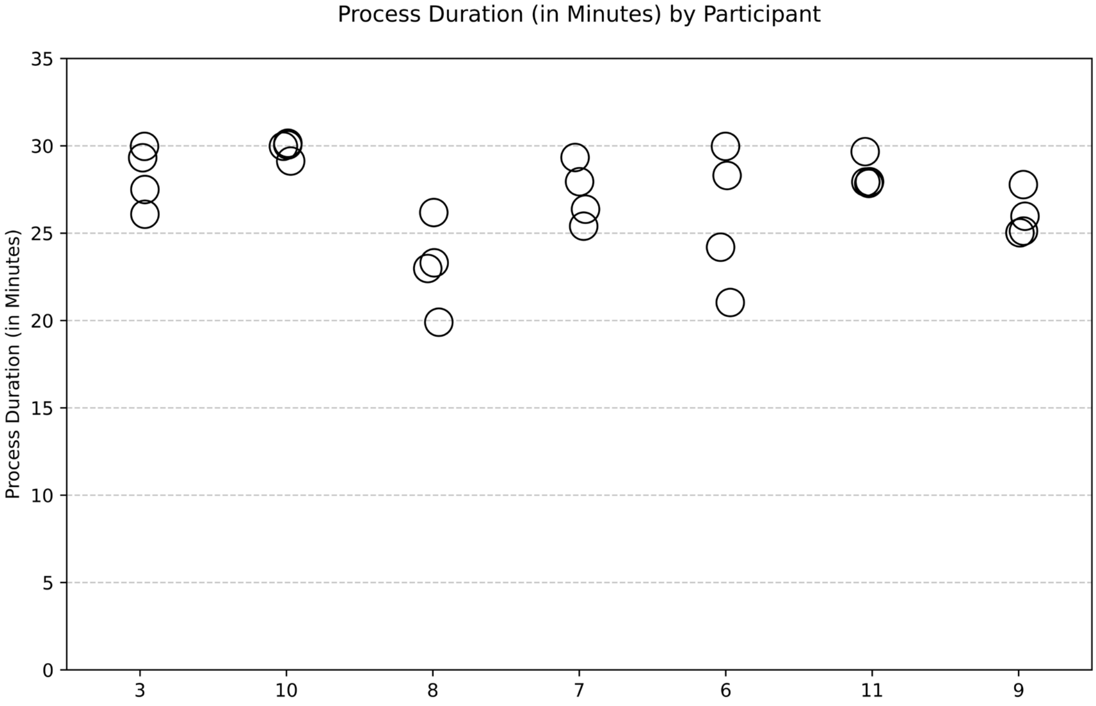

Figure 2. Plot showing the four process durations of each participant. Participants are ordered from low to high based on their FUT B scores.

The statistics for tool interaction frequency and percentage of tool use are reported in Table 4. As shown in the table, participants consulted tools at least once per minute while writing, and these consultations accounted for at least one-sixth of the writing process.

Table 4. Descriptive tool interaction statistics (*N* = 7)

Statistic	Tool consultation frequency (per minute)	Percentage of tool use
M	1.10	21.49
SD	0.22	6.36
95% CI	[0.90, 1.31]	[15.61, 27.37]
Range	0.94–1.56	16.63–35.05

During the post-task interviews, all participants stated that they perceived the writing tasks as suitable for their proficiency. Participants 6, 8, and 10 explained that this was because the topics had been covered in class. Participants 6 and 8 also gave as an additional reason that the writing tasks were proficiency-independent. Participant 8 attributed this to the open-ended nature of the writing tasks. However, alongside their general satisfaction, some participants also expressed concerns. For example, Participant 9 noted that the topics may be unfamiliar to those who have never visited the Nordics before. Participant 3, a first-year student, remarked that they had to adjust to the writing tasks, as they were not used to writing in Swedish without in-class pre-writing activities that provided relevant vocabulary, grammar, and topic knowledge. Similarly, Participant 10, another first-year student, mentioned that writing half a page was not feasible for them because of their limited L2 knowledge.

Discussion

L2 Swedish Proficiency Tests

Our first objective was to analyse the suitability of the FUT B and the SweLT to estimate the L2 Swedish proficiency of our participants. Of the four tests we identified, these two emerged as the most appropriate for this purpose. They appeared shortest in terms of number of items, while still targeting dimensions of proficiency that are relevant in L2 writing. In addition, they both yield a single, granular score.

The strong positive correlation between participants' scores on the SweLT and FUT B indicates that the two proficiency tests ranked our participants similarly. Moreover, no participants obtained scores near the minimum or maximum possible score on either the FUT B or the SweLT, which signifies that both tests provided opportunities for both lower- and higher-proficiency participants to demonstrate their actual proficiency. However, the SweLT and the FUT B also differ in several ways. Participants' mean score of just above 50% on the FUT B indicates that this test was of moderate difficulty, i.e., that the test items were neither too easy nor too hard for the average participant. With an average score of approximately 70%, the SweLT appeared less challenging. The two proficiency tests also differ in test format: the FUT B incorporates open-ended questions, whereas the SweLT is composed solely of multiple-choice questions, which can be more susceptible to correct answers through guessing. Furthermore, its larger standard deviation and wider range indicates that the FUT B can capture more subtle differences among our participants. Lastly, the FUT B is notably quicker to administer. In experimental writing studies, participants are often tasked with multiple writing tasks, placing already heavy demands on their cognitive resources. Assuming that faster tests are less burdensome, the FUT B may be more suited to estimate L2 proficiency in this context.

Correlations between participants' scores on the two proficiency tests and two other proficiency measures – that is, the year of L2 Swedish instruction in which they were enrolled and their self-assessments – also tended to be positive, although slightly weaker in strength. Our qualitative analysis showed that deviating test scores given participants' years of instruction could be attributed to several factors: longer test completion times, time spent in Sweden, varying skills within the same year, and the coarse classification by full years of instruction. The deviations between the proficiency test scores and the self-assessments might be due to undue modesty or inflated self-esteem (Wall et al., 1994), or to participants comparing their skills to their classmates rather than assessing them on a broader scale. These findings further illustrate the importance of using independent tests rather than relying on year of instruction or self-assessments to estimate L2 proficiency in SLA research, as highlighted by Park et al. (2022).

Writing Tasks

Our second objective was to design a set of writing tasks and report on their suitability to elicit suitable writing product and process data (including tool interactions) from our participants. In addition, we aimed to assess the extent to which our participants perceived the writing tasks as suitable for their proficiency.

The text length statistics revealed that even the participant with the shortest text lengths produced more than 100 words on average, meeting the minimum threshold recommended for various indices commonly used to study L2 writing products (Hwang & Polio, 2023; Zenker & Kyle, 2021). This suggests that the writing tasks elicited suitable data for investigating the L2 writing products of all participants, regardless of their proficiency.

To increase the comparability of participants' writing processes, we implemented a 30-minute time limit per writing task. Participants' process durations indicated that their writing processes often approached but rarely lasted the full 30 minutes. Moreover, even when participants did use the entire 30 minutes, both the product and interview data suggest that they completed their writing process by the end of the limit. These findings indicate that the writing tasks allowed for capturing full writing processes across proficiencies, without cutting them short.

The statistics also showed that all participants managed to produce fairly balanced texts, containing both descriptive and expository writing, even though the latter is usually expected only from more proficient learners (Council of Europe, 2020). This suggests that, despite incorporating genres of varying difficulty, the writing tasks elicited comparable products and processes across proficiencies.

Participants' tool consultation frequency and percentage of time spent in tools implied that the writing tasks prompted even the higher-proficiency participants to regularly consult tools during their writing process. This suggests that the writing tasks elicited data suited to investigate how tool use affects the L2 writing of our participants across proficiencies.

Lastly, the post-task interviews indicated that all participants perceived the writing tasks as suitable for their proficiency, with some attributing this to the writing tasks being proficiency-independent and some to their familiarity with the topics. However, some responses also contained concerns about not meeting the text length target and the difficulty of the writing tasks when lacking relevant vocabulary, grammar, and topic knowledge. To address these concerns, two modifications could be made in a future version of the writing tasks. First, since the descriptive statistics indicate that even the participant with the longest text lengths did not reach the 400-word target, this target could be lowered to slightly above the mean text length produced in this study (approximately 300 words). This revised target would also align better with the targets used in other studies with participants of varying proficiency. Second, while providing participants with L2 vocabulary and grammatical structures is not feasible as it may distort the writing products, a future version could provide more pictorial input. Multiple pictures per subtopic, with each picture highlighting a different aspect of the subtopic, could help to fill any potential gaps in topic knowledge.

Concluding Remarks

We conducted a preliminary study to address two needs identified during instrument selection for a research project on L2 writing involving learners of Swedish of varying proficiency. First, we needed a test to estimate participants' L2 Swedish proficiency, but deciding which one to use proved difficult because no studies exist that compare the available tests. We therefore analysed the two tests that seemed most appropriate as proficiency estimators for experimental studies on L2 Swedish (writing). Our findings revealed that the FUT B is more efficient, as it performs similarly to the SweLT but is quicker to administer.

Second, we needed a set of writing tasks suited to collect writing data from learners of varying proficiency. However, we observed that other L2 writing studies with participants of varying proficiency rarely discuss whether their writing tasks can elicit suitable data for analysis across proficiencies. We therefore developed our own writing tasks and investigated their ability to elicit such data using five measures that focused on different aspects of the writing products and processes, along with interview data on participants' perceptions of the writing tasks. Overall, the writing tasks performed well. However, the interview data also indicated that they could be further refined by adjusting the text length recommendations and supplying participants with multiple pictures per subtopic.

Due to the small population of L1 Dutch learners of Swedish, the sample size for this study is relatively limited. To mitigate this limitation, we collected extensive data from each participant, including four L2 proficiency estimates, product and process data from four writing tasks, and interview data on their perceptions of the writing tasks. This range of data allowed us to address the research objectives from multiple perspectives.

Future studies could build on this study in several ways. For example, they could examine the correlation between the proficiency test scores and L2 writing skills to verify our assumption that the FUT B and SweLT measure dimensions of L2 proficiency particularly relevant to L2 writing. Additionally, data on learners' perceptions of the proficiency tests could be gathered to further support our assumption that the FUT B, being quicker to administer than the SweLT, is the least burdensome of the two. Finally, expert evaluations of the writing tasks, similar to those in Lu (2020) and Alanen et al. (2010), could provide an additional perspective on their suitability for use across proficiencies.

References

Alanen, R., Huhta, A., & Tarnanen, M. (2010). Designing and assessing L2 writing tasks across CEFR proficiency levels. In I. Bartning, M. Martin, & I. Vedder (Eds.), *Communicative proficiency and linguistic development* (pp. 21–56). European Second Language Association. https://eurosla.org/monographs/EM01/21-56Alanen_et_al.pdf

Barkaoui, K. (2019). What can L2 writers' pausing behavior tell us about their L2 writing processes? *Studies in Second Language Acquisition*, *41*(3), 529–554. https://doi.org/10.1017/S027226311900010X

Barrot, J. S., & Agdeppa, J. Y. (2021). Complexity, accuracy, and fluency as indices of college-level L2 writers' proficiency. *Assessing Writing*, *47*, Article 100510. https://doi.org/10.1016/j.asw.2020.100510

Bokander, L. (2016). SweLT 1.0: Konstruktion och pilottestning av ett nytt svenskt frekvensbaserat ordförrådstest. *Nordand: Nordic Journal of Second Language Research*, *11*(1), 9–30.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (Vol. 32, pp. 21–46). John Benjamins Pub. Co. https://doi.org/10.1075/lllt.32

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing. https://www.coe.int/en/web/common-european-framework-reference-languages

Coyle, Y., Nicolás-Conesa, F., & Cerezo, L. (2023). Overview of methodological procedures in research on written corrective feedback processing. In R. M. Manchón & J. Roca De Larios (Eds.), *Research Methods in the Study of L2 Writing Processes* (Vol. 5). John Benjamins Pub. Co. https://doi.org/10.1075/rmal.5

Długosz, K. (2023). Cross-linguistic influence in the comprehension of reflexive possessive pronouns in L3 and L4 Swedish. *International Journal of Bilingualism*, Advance online publication. https://doi.org/10.1177/13670069231194341

Forsbom, E. (2006). *Deriving a base vocabulary pool from the Stockholm-Umeå Corpus*. Term paper for NGSLT course Soft Computing.

Gánem-Gutiérrez, G. A., & Gilmore, A. (2018). Tracking the real-time evolution of a writing event: Second language writers at different proficiency levels. *Language Learning*, *68*(2), 469–506. https://doi.org/10.1111/lang.12280

Huhta, A., Luoma, S., Oscarson, M., Sajavaara, K., Takala, S., & Taesdale, A. (2002). DIALANG: A diagnostic language assessment system for adult learners. In *Common European framework of reference for languages: Learning, teaching, assessment. Case studies.* (pp. 130–145). Council of Europe Publishing. https://rm.coe.int/168069f403

Hwang, H.-B., & Polio, C. (2023). Text length effects on the reliability of syntactic complexity indices. *Research Methods in Applied Linguistics*, *2*(3), Article 100085. https://doi.org/10.1016/j.rmal.2023.100085

Källgren, G., Gustafson-Capková, S., & Hartmann, B. (2006). *Manual of the Stockholm Umeå Corpus version 2.0*. Stockholm University. https://spraakbanken.gu.se/parole/Docs/SUC2.0-manual.pdf

Khushik, G. A., & Huhta, A. (2020). Investigating syntactic complexity in EFL learners' writing across Common European Framework of Reference levels A1, A2, and B1. *Applied Linguistics*, *41*(4), 506–532. https://doi.org/10.1093/applin/amy064

Kim, Y., Nam, J., & Lee, S.-Y. (2016). Correlation of proficiency with complexity, accuracy, and fluency in spoken and written production: Evidence from L2 Korean. *Journal of the National Council of Less Commonly Taught Languages*, *19*, 147–181.

Kojima, M., Kaneta, T., & In'nami, Y. (2022). L2 writing and its external correlates. In E. H. Jeon & Y. In'nami (Eds.), *Understanding L2 Proficiency: Theoretical and meta-analytic investigations* (Vol. 13, p. 388). John Benjamins Pub. Co. https://doi.org/10.1075/bpa.13.06koj

Kostromitina, M., & Plonsky, L. (2022). Elicited imitation tasks as a measure of L2 proficiency: A meta-analysis. *Studies in Second Language Acquisition*, *44*(3), 886–911. https://doi.org/10.1017/S0272263121000395

Lahuerta, A. C. (2018). Study of accuracy and grammatical complexity in EFL writing. *International Journal of English Studies*, *18*(1), 71–89. https://doi.org/10.6018/ijes/2018/1/258971

Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication*, *30*(3), 358–392. https://doi.org/10.1177/0741088313491692

Liao, J. (2022). The effects of linguistic measures in the analysis of L2 Chinese descriptive writing. *Chinese as a Second Language*, *56*(2), 118–145. https://doi.org/10.1075/csl.21012.lia

Lindberg, I., & Johansson, S. (2018). The development of Swedish receptive vocabulary in CLIL: A multilingual perspective. In L. K. Sylvén (Ed.), *Investigating content and language integrated learning* (pp. 236–258). Multilingual Matters. https://doi.org/10.21832/9781788922425-019

Lu, X. (2020). *Writing in a non-alphabetic language using a keyboard: Behaviours, cognitive activities and text quality* [Doctoral dissertation, University College London]. https://discovery.ucl.ac.uk/id/eprint/10106308/9/Lu_10106308_thesis.pdf

Milton, J. (2013). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis* (pp. 57–78). European Second Language Association. https://www.eurosla.org/monographs/EM02/Milton.pdf

Park, H. I., Solon, M., Dehghan-Chaleshtori, M., & Ghanbar, H. (2022). Proficiency reporting practices in research on second language acquisition: Have we made any progress? *Language Learning*, *72*(1), 198–236. https://doi.org/10.1111/lang.12475

Polio, C., & Friedman, D. (2016). *Understanding, Evaluating, and Conducting Second Language Writing Research*. Routledge. https://doi.org/10.4324/9781315747293

Prentice, J., & Forsberg Lundell, F. (2021). Productive collocation knowledge and advanced CEFR-levels in Swedish as a second language: A conceptual replication of Forsberg Lundell, Lindqvist & Edmonds (2018). *Journal of the European Second Language Association*, *5*(1), Article 1. https://doi.org/10.22599/jesla.72

Révész, A., Michel, M., Lu, X., Kourtali, N., Lee, M., & Borges, L. (2022). The relationship of proficiency to speed fluency, pausing, and eye-gaze behaviours in L2 writing. *Journal of Second Language Writing*, *58*, Article 100927. https://doi.org/10.1016/j.jslw.2022.100927

Ribeck, J., Jansson, H., & Sköldberg, E. (2014). Från *aspekt* till övergripande: En ordlista över svensk akademisk vokabulär. *Nordiske studier i leksikografi*, *12*, 370–384.

Sasaki, M. (2000). Toward an empirical model of EFL writing processes: An exploratory study. *Journal of Second Language Writing*, *9*(3), 259–291. https://doi.org/10.1016/S1060-3743(00)00028-X

Sasayama, S., Garcia Gomez, P., & Norris, J. M. (2021). *Designing efficient L2 writing assessment tasks for low-proficiency learners of English* (pp. 1–31) [TOEFL Research Report 97 and ETS Research Report 21-27]. Educational Testing Service. https://doi.org/10.1002/ets2.12341

Spelman Miller, K., Lindgren, E., & Sullivan, K. P. H. (2008). The psycholinguistic dimension in second language writing: Opportunities for research and pedagogy using computer keystroke logging. *TESOL Quarterly*, *42*(3), 433–454. https://doi.org/10.1002/j.1545-7249.2008.tb00140.x

Sun, Y., & Lan, G. (2023). A bibliometric analysis on L2 writing in the first 20 years of the 21st century: Research impacts and research trends. *Journal of Second Language Writing*, *59*, Article 100963. https://doi.org/10.1016/j.jslw.2023.100963

Tiryakioglu, G., Peters, E., & Verschaffel, L. (2019). The effect of L2 proficiency level on composing processes of EFL learners: Data from keystroke loggings, think alouds and questionnaires. In E. Lindgren & K. P. H. Sullivan (Eds.), *Observing writing* (pp. 212–235). Brill. https://doi.org/10.1163/9789004392526_011

Tiselius, E., & Sneed, K. (2020). Gaze and eye movement in dialogue interpreting: An eye-tracking study. *Bilingualism: Language and Cognition*, *23*(4), 780–787. https://doi.org/10.1017/S1366728920000309

Vallejos, C. A. (2020). *Fluency, working memory and second language proficiency in multicompetent writers* [Doctoral dissertation, Georgetown University]. http://hdl.handle.net/10822/1059667

Vasylets, O., & Marín, J. (2021). The effects of working memory and L2 proficiency on L2 writing. *Journal of Second Language Writing*, *52*, Article 100786. https://doi.org/10.1016/j.jslw.2020.100786

Wall, D., Clapham, C., & Alderson, J. C. (1994). Evaluating a placement test. *Language Testing*, *11*(3), 321–344. https://doi.org/10.1177/026553229401100305

Weigle, S. C. (2010). *Assessing writing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732997

Xu, C., & Ding, Y. (2014). An Exploratory Study of Pauses in Computer-Assisted EFL Writing. *Language Learning & Technology*, *18*(3), 80–96. http://dx.doi.org/10125/44385

Yoon, H.-J., & Polio, C. (2017). The Linguistic Development of Students of English as a Second Language in Two Written Genres. *TESOL Quarterly*, *51*(2), 275–301. https://doi.org/10.1002/tesq.296

Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing*, *47*, Article 100505. https://doi.org/10.1016/j.asw.2020.100505

Zhang, X. (2013). The *I don't know* option in the vocabulary size test. *TESOL Quarterly*, *47*(4), 790–811. https://doi.org/10.1002/tesq.98

Notes

[1] https://dialangweb.lancaster.ac.uk/

[2] See https://osf.io/a2xrf for PDF versions of the two proficiency tests and links to create direct copies of the tests in Google Forms.

[3] An example of duplicate words would be the three occurrences of *likadan* (similar). An example of near duplicate words would be *osäker* (uncertain) and *osäkerhet* (uncertainty). In total, we replaced seven words with alternatives. To qualify as an alternative, a word had to be listed within the same frequency level in Forsbom's (2006) vocabulary pool and had to belong to the same part of speech.

[4] https://www.inputlog.net/

Appendix

The instructions for the 'workshops' writing task in the dictionary condition, translated from Dutch into English by the first author. The instructions for the other writing tasks and the machine translation condition only differ minimally from this example. The complete set of instructions, along with the accompanying pictures, are available on request from the first author of this article.

READ THESE INSTRUCTIONS CAREFULLY

- **TASK:** Write a text in Word with two parts:
    1. In the first part, you ***describe*** and ***interpret the pictures***. What do you see and what do you think is happening?
    2. In the second part, you think about which of the three workshops you find the most appealing: Nordic knitting (picture 1), ice sculpting (picture 2) or baking cinnamon rolls (picture 3)? ***Also explain*** why.

- **LANGUAGE**: Write your text in ***Swedish***. Try to write a text that is ***as polished as possible***.
- **TIME**: You have a maximum of ***30 minutes*** for this task. If you think your text is finished, you can hand it in before the 30 minutes are up. The timer at the bottom right of your screen will show you how much time you have left.
- **LENGTH**: Try to write a text about ***half a page*** long. Make sure that the description of the pictures (first part) and the part in which you explain your preference (second part) are about ***the same length***.
- **TOOLS**: While writing, you are (ONLY) allowed to consult the ***Swedish-Dutch*** and ***Dutch-Swedish*** dictionaries on ***the Van Dale website***. Do NOT consult other dictionaries. Do NOT use Word's spelling and grammar checker.
- **PROCEDURE**: Word and Van Dale are already open. Do not open any other tabs or windows in Chrome. Do not place Chrome and Word side by side by splitting your screen, but instead switch between the two programmes via the taskbar. Do not zoom in or out either. Please let me know when you are finished, and I will come to close your document.