Assessing L2 English Skills in an Online Environment: What Can This Look Like and How to Assess L2 English Writing Skills?

Vanessa De Wilde, Geert De Meyer, and Pedro De Bruyckere

Abstract Studies with young L2 English learners have shown differences in learners' L2 English proficiency. This creates a situation in which learners in a classroom often form a very heterogeneous group. In this study, we report on the development of an online tool to assess pupils' proficiency level at the start of formal instruction in Flanders. We used group concept mapping and a teacher questionnaire to investigate what this tool should look like. Results indicated that teachers opted for an online test that contained tasks for all four language skills. In the second part of the chapter, we report on one of the challenges that came with the development of this online tool, i.e. finding a method to assess learners' writing that is both reliable and easy to use. In order to do this, we explored the possibility of using benchmark texts which were selected in a two-stage approach using comparative judgment. Results showed that this method with five benchmark texts that teachers can use to correct their learners' writing can indeed be used reliably and efficiently.

Keywords Benchmark texts · Writing · Assessment · Online assessment · Adolescent learners · Evaluation

V. De Wilde (✉)
Department of Translation, Interpreting and Communication, Ghent University, Ghent, Belgium

Artevelde University of Applied Sciences, Ghent, Belgium
e-mail: vanessa.dewilde@ugent.be

G. De Meyer
Department of Linguistics, Ghent University, Ghent, Belgium

Artevelde University of Applied Sciences, Ghent, Belgium
e-mail: geert.demeyer@ugent.be

P. De Bruyckere
Utrecht University, Utrecht, The Netherlands

Artevelde University of Applied Sciences, Ghent, Belgium
e-mail: p.debruyckere@uu.nl

M.-M. Suárez, W. M. El-Henawy (eds.), *Optimizing Online English Language Learning and Teaching*, English Language Education 31,
https://doi.org/10.1007/978-3-031-27825-9_16

1 Introduction

This chapter reports on the development of an online tool designed to measure L2 English learners' proficiency at the CEFR A2-level and as such inform teachers about the differences in these learners' proficiency levels. The chapter will first discuss the development of the online tool which started from the needs expressed by Flemish L2 English teachers. Secondly, the chapter will focus on a valid and reliable way to assess learners' writing skills. In order to be able to assess the learners' writing in a reliable yet time-efficient manner, we explored the possibilities of a two-stage approach for marking L2 English writing using comparative judgment and benchmark texts.

Below we will discuss the need for an online tool measuring L2 English proficiency, some of the difficulties concerning the assessment of writing skills and the context in which this study took place. We will then describe the different steps that were taken in the development of the online tool, discuss the results of the study and end with some pedagogical implications.

2 Literature Review

2.1 The Need for an Online Assessment Tool

Studies with primary school age L2 English learners have found considerable differences in L2 English proficiency even before the start of formal L2 English instruction (De Wilde et al., 2020; Muñoz et al., 2018; Puimège & Peters, 2019). These large differences in learners' prior L2 English knowledge pose considerable challenges to teachers so knowing about these differences is important as prior knowledge can have a huge impact on further learning, e.g. in relation to the amount of instruction that is needed (e.g. De Bruyckere, 2017; Hattie & Yates, 2013). Therefore, it was decided to develop a test to give teachers an opportunity to get information about their learners' L2 English proficiency level at the start of the lessons in secondary school. The test is meant to inform teachers so they can adapt lessons to the learners' various levels of proficiency and prior knowledge of English (e.g. through differentiated instruction). It is thus meant to be a low-stakes test.

2.2 Assessing Writing

Assessing students' writing comprises many different aspects such as content, organization, and linguistic features. Therefore, scoring writing tasks is often considered a challenging and time-consuming exercise (Hamp-Lyons, 1990; Schoonen, 2005). Teachers and researchers have studied many different methods to rate writing skills.

A distinction is often made between analytic and holistic methods. In order to rate writing tasks analytically, raters often use rubrics that list criteria that should be taken into account often also containing descriptors of the expected performance for the different levels of each criterion. In an analytic scoring method, the final score is a combination of partial scores (Crusan, 2013). Holistic scoring methods look at the text as a whole and attribute one single score to the writing product, whereas analytic scoring methods give different scores for different aspects of the text, such as linguistic or content-related aspects (Harsch & Martin, 2013). When scoring a writing task in a holistic manner, raters sometimes use a set of criteria that need to be considered when rating the task, but these criteria serve as a guideline to give one overall score. There are also other methods to assess writing tasks holistically. Two of these methods will be discussed below.

2.2.1 Two Holistic, Comparative Approaches: Comparative Judgment and Benchmark Texts

Apart from the more traditional analytic and holistic approaches in which students' writing is assessed in an absolute manner, there are also comparative approaches, in which representations (e.g. written texts or images) are compared.

A method that has recently received some attention is comparative judgment, inspired by the work of Thurstone (1927), who claimed that people's judgment is more reliable when comparing performances than when judging a single performance. The method was introduced into education by Pollitt and Murray (1996). In this method, multiple raters compare pairs of representations (e.g. written texts) and decide which of the two representations is the better one. After the raters have made a set number of comparisons of all the tasks, each learner's writing task is assigned a place on a rank order ranging from the weakest to the strongest. The overall quality of a writing task is thus based on repeated comparisons (Lesterhuis et al., 2017). Recently, research teams have set up studies to organize this type of rating process digitally. To build an information system that could be used for comparative judgment, Coenen et al. (2018) identified several design requirements for the tool to be a success. These were: being able to do valid and time-efficient assessments, reduce cognitive load, increase reliability, support competence development, and support accountability. The tool which resulted from this study is called Comproved (www.comproved.com) but similar tools are available (e.g. No More Marking, Jones, 2016). The studies mentioned above have shown that this approach can result in reliable ratings, but various raters are needed, and many comparisons have to be made. The guidelines on the Comproved website for example, mention that for a reliability of .70 the following formula should be used: *number of representations * 7.5 / number of raters = number of comparisons per rater.* This shows that the procedure can be quite time-consuming which might be a hindrance for teachers in day-to-day classroom practice, as the number of holistic comparisons to be done can be high (Humphry & Heldsinger, 2020; Lesterhuis et al., 2017).

Another form of holistic rating can be done with the use of benchmark texts (Lesterhuis et al., 2017). In this procedure, several texts are chosen which represent different levels of overall writing quality to serve as benchmarks. Teachers, or other raters, then compare their students' work with chosen benchmarks and decide which of the texts resembles their students' work the most. The level associated with the most suitable benchmark text is the level allocated to their students' writing. Bouwer et al. (2016) investigated possibilities of rating written texts with benchmark texts and found that benchmark ratings were as reliable as 'absolute' analytic and holistic ratings. They did this on paper however, while the aim of this study is to check if this can also be a suitable approach online.

Recently, several studies which investigated L1 writing have adopted an integrated, two-stage approach that combines the use of comparative judgment and benchmark texts (De Smedt et al., 2020; Humphry & Heldsinger, 2020; McGrane et al., 2018). In this approach, first, a set of written texts are compared through comparative judgment. After this procedure, experts choose a number of texts from this set that represent different levels of writing quality as benchmark texts. These benchmarks are then used when scoring new, similar writing tasks. Below, we report on a study in which we have adopted this two-stage approach for an L2 picture narration task to investigate whether using this approach which has been used in L1 writing, is also appropriate in L2 writing in an online environment.

This chapter reports on the development of a tool meant to assess L2 English learners' proficiency level. It describes the process towards a test structure and content that meets teachers' needs. It further investigates a method to address specific challenges concerning the assessment of written texts. The following research questions are central in this chapter:

RQ1: What should an online tool which aims to map learners' L2 English proficiency at the start of formal instruction look like?

RQ2: How efficient and reliable is the assessment of L2 English writing tasks following a two-stage approach (combining the use of comparative judgement and benchmark texts)?

3 Context of the Study

Formal L2 English lessons are compulsory in Flanders, the Northern part of Belgium, from the first or second year of secondary school onwards, when learners are 12–14 years old. The starting age for English is rather late when compared to other European countries (Enever, 2011) because English is the second foreign language to be taught in Flemish schools. The first foreign language which is taught in Flanders is French, which is an official language in Belgium.

Pupils are expected to reach the A2 level of the Common European Framework of Reference (Council of Europe, 2009) for English by the end of the second year of secondary school. In primary education, L2 English is not a compulsory part of the curriculum, and formal instruction only starts in secondary school. However, this does not mean English is absent in most learners' daily lives. Most learners have been exposed to English regularly before the start of the lessons (for example, when gaming or watching television), and this leads to big differences in pupils' prior knowledge of English. De Wilde et al. (2020) did a study with 780 Flemish learners who were in the last year of primary school. They found that 25% of Flemish 11-year-olds obtained a score of 80% or higher on an A2-level listening test with a mean of 15/25 but overall, there was a broad range of test scores (from 0 to 25 out of 25). For the A2-level speaking task, scores were considerably lower (with a mean of 7/20), but still, a considerable number of the participants scored 80% or higher (14% of the participants). Finally, 10% of Flemish 11-year-olds obtained a score of 80% or higher on an A2-level reading and writing test, whereas more than half of the participants obtained a test score below 50%, again pointing to large individual differences prior to the start of formal instruction.

The online tool presented in this chapter was developed to give teachers in Flanders more insight in the actual differences in their learners' L2 English proficiency level. First, we investigated what teachers expected from such a tool and in a second study we looked into an efficient and reliable way to score learners' writing.

4 Research Question 1

Following, we proceed to explain the methodology and results obtained to answer our first research question, which is: "What should an online tool which aims to map learners' prior knowledge look like?"

4.1 Methodology

4.1.1 Instruments and Procedure

To be able to develop a test that would meet teachers' needs, we decided to consult teachers and other stakeholders before the actual test development was started. The teacher questionnaire was designed using group concept mapping (GCM). This method, which was developed by Trochim (1989a, b) and further adapted by Stoyanov and Kirchner (2004), can be used to gather and organize ideas in a structured manner. In this study, it was used in a simplified version which consisted of three rounds. In round one, we sent a list with open questions to experts in the field of education and assessment to gather answers which could lead to items for the questionnaire. The open questions were listed in an online form, the link to the form was then e-mailed to the experts who answered the questions anonymously. They were given 1 week to answer the questions. In round two, we sent the same seven experts a set of possible items for the questionnaire, which were based on the answers to the open questions from the first round. We asked the experts to rate how important these items were. Again, they had 1 week to complete the questionnaire. In the last round, we made an initial version of the questionnaire for the teachers and asked a focus group with five new participants with expertise in language testing and/or foreign language teaching to comment on the questionnaire and give suggestions for improvement. After the focus group, we made a second and final version of the questionnaire, which we made available for teachers. The questionnaire consisted of some questions asking about their teaching and experience and a number of statements about what they thought an L2 English test for their learners should look like. Answers to the statements were given on a Likert scale ranging from 1 (totally unimportant) to 5 (very important). The questionnaire can be found in Appendix A.

4.1.2 Participants

As mentioned above, we consulted teachers and other stakeholders in order to be able to have a clear view on their expectations for an L2 English proficiency test. In the first phase, we consulted experts in the field of foreign language education and assessment such as scholars, policymakers, and curriculum designers (n = 12). Seven experts took part in the group concept mapping procedure and five experts took part in the focus group. The participants in the focus group were part of the advisory committee for this research project.

In a second phase, 95 participants filled in the teacher questionnaire. Most participants were teachers in the first 2 years of secondary school (n = 64), 29 teachers also taught in secondary school but taught older pupils, one participant was a teacher trainer and one educational adviser for English took part in the study. The teachers who completed the questionnaire had various degrees of experience (between 1 and over 30 years of experience).

4.1.3 Analysis

The results of the teacher questionnaire were analyzed quantitatively and used to make decisions about the structure, content, and duration of the test. Descriptive statistics can be found in the results section.

4.2 Results

Teachers' answers showed that they believed a test should contain activities looking into learners' prior knowledge of the language skills (cf. Table 1). Therefore, it was decided that the test should consist of four parts, each testing one language skill: listening comprehension, reading comprehension, writing, and speaking.

Table 1 Teacher questionnaire: descriptive statistics test characteristics (Likert scale 1–5)

This test should…	Min	Max	Mean	SD
Focus on realistic and real language.	1	5	4.53	0.79
Focus on academic language.	2	5	3.65	0.78
Focus on productive language.	1	5	4.39	0.89
Focus on receptive language.	1	5	4.22	0.71
Be linked to the CEFR.	2	5	3.96	0.83
Measure the four skills.	1	5	4.53	0.85
Measure listening skills.	1	5	4.42	0.72
Measure reading skills.	1	5	4.43	0.66
Measure speaking skills.	2	5	4.45	0.68
Measure writing skills.	2	5	4.21	0.77
Measure grammatical knowledge.	1	5	4.01	0.89
Measure lexical knowledge.	1	5	4.23	0.81
Give feedback to the pupils.	1	5	4.45	0.80
Give teachers the opportunity to give feedback to the pupils.	1	5	4.58	0.74
Be done on paper.	1	5	2.79	1.01
Be done in a digital manner.	1	5	3.26	1.04

Another important aspect for teachers was the possibility to give feedback. The form of the test was less important for the teachers than the content, but they seemed to favor a digital test over a paper-based test. Descriptives statistics of the scores given by the teachers can be found in Table 1.

The majority of the teachers (60%) also asked that the duration of the test would be approximately 50 min, the equivalent of one teaching period in Flanders, 31% of the teachers opted for a shorter test (30 min) and 9% of the teachers would also use a test which would take more than 50 min. We decided to settle for a 50-min test. As the test is meant for learners who are at the start of formal education and learners could be absolute beginners, the instructions had to be available in both English and Dutch, which is the language of instruction.

During test development, we considered the test's practicality, and we decided a type of scoring was needed that would be easy to use for the teachers, as they would be the ones scoring their learners' tests. For the scoring of the writing task, we decided to investigate the possibility of using benchmark texts which were selected via a two-stage approach. This procedure will be discussed below.

5 Research Question 2

After having analyzed the testing tool, we proceed to answer our second research question, which is: "How efficient and reliable is the assessment of L2 English writing tasks using benchmark texts which are selected via a two-stage approach?". We will do so through two different studies.

5.1 *Study 1*

5.1.1 Methodology

Participants

In order to be able to answer the second research question, 121 participants wrote one or two written texts. All the participants were at the start of formal L2 English education and were between 12 and 14 years old. We tested pupils in six classes in two different schools, three classes per school. The participants from school one were in the first year of secondary school, those from school two were in the second year of secondary school. All participants had just started formal L2 English education. They had received less than 15 h of formal English instruction.

Fifty-three raters took part in the comparative judgment procedure. All the raters had experience with rating L2 writing tasks: They were either working as teachers or teacher trainers (n = 11) or they were in the second year of a three-year bachelor's program in which they were trained as English teachers (n = 42). The students from the teacher training program had already done a teaching practice in a secondary school and had been trained to score students' work.

Instruments and Procedure

To be able to capture the different levels of proficiency among the learners while still giving sufficient support to the true beginners, we decided to use a picture narration task. According to Goodier and Szabo, the authors of the Collated Representative Samples of Descriptors of Language Competencies Developed for Young Learners (2018), the task of telling a simple story is a relevant task at the CEFR A2-level for learners aged 11–15 years. The visuals that were added in the writing task in the present study were meant to give extra support to learners with a low(er) proficiency level. Three different picture stories were designed. An example of one of the picture stories can be found in Fig. 1 below.

In the first phase of the study, we collected 177 writing samples. The participants described a set of four pictures which together made up a story. There were three different stories (picture story A, n = 60; picture story B, n = 56; picture story C, n = 61). The picture stories were designed in such a manner that all pupils would be able to relate to the situations depicted in the stories. No explicit time limit was given to the pupils. The writing tasks were paper-based and were digitalized by the researchers for the next phase of the study (comparative judgment).

The learners' writing tasks were rated using the comparative judgment tool Comproved (Coenen et al., 2018). In this tool, raters are asked to compare two representations, in this case two of the 177 texts, and to decide which of the

(continued)

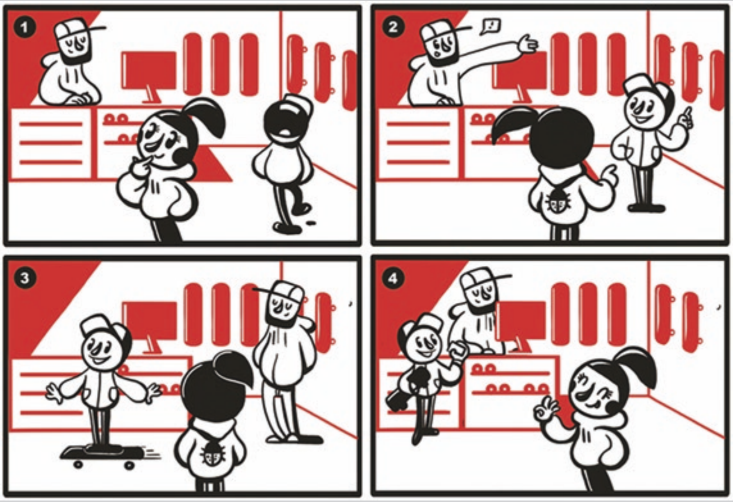


Fig. 1 Example of a picture story designed for the picture narration task

two representations is the better one. There were 53 raters who each made 33 comparisons, resulting in a total of 1749 comparisons. The number of comparisons is sufficient in order to obtain reliable results (cf. formula: number of representations * 7.5 / number of raters = number of comparisons per rater). Per comparison, raters were asked to select the best representation of two. There were no further instructions concerning how they should rate the writing sample, no criteria were given for the assessment. They were only asked to indicate which writing sample they believed had the highest quality of the two samples they were presented with in each comparison. Raters could choose to add some comments to justify their decision, but this was not obligatory, and it was not taken into account when making the rank order. Following the procedure of the two-stage approach (Humphry & Heldsinger, 2020), the results of the comparative judgments procedure were used to inform the choice of the benchmark texts. Descriptors of the benchmark texts were taken from the CEFR descriptors for young learners aged 11–15 years (Goodier & Szabo, 2018).

Analysis

Descriptive statistics of the rank order of the 177 representations (written texts) that resulted from the comparative judgment procedure are given in the results section. To investigate the reliability of the rank order, the scale separation reliability (SSR) was calculated. Verhavert et al. (2018) found this measure (with values between 0 and 1) can be used as an index for interrater reliability in comparative judgment.

5.1.2 Results

The raters' work resulted in a reliable rank order of the 177 representations. In the current study, the SSR was high (.88), indicating that there was strong agreement between the raters on the quality of the written texts. Thus, we could be confident that the rank order that followed the 1749 comparisons was reliable.

We then compared the results of the comparative judgment procedure for the three different picture stories and chose the picture story with the best spread in results. The boxplot in Fig. 2 shows the spread of the scores of each representation (written text) per picture story. In Table 2, the descriptive statistics for the results of the comparative judgment procedure for the representations per picture story are given. Figure 2 and Table 2 show that the spread in the representations for story B (which is the example story given in Fig. 1) is almost evenly divided. The mean ability is around zero, some representations received a high score (maximum score = 5.52), others have a very low score (minimum score: −5.87), there are no outliers. We, therefore, decided to continue the study with story B in stage two.

After the comparative judgment procedure, two researchers selected four benchmark texts. The selection was based on the rank order of the representations which was decided by the 53 raters who took part in the comparative judgment procedure. Starting from that order, the researchers chose texts which were a good representation of the four different levels they wanted to discriminate: above A2, A2, A1 and below A1 based on the level descriptors found in the Common European Framework

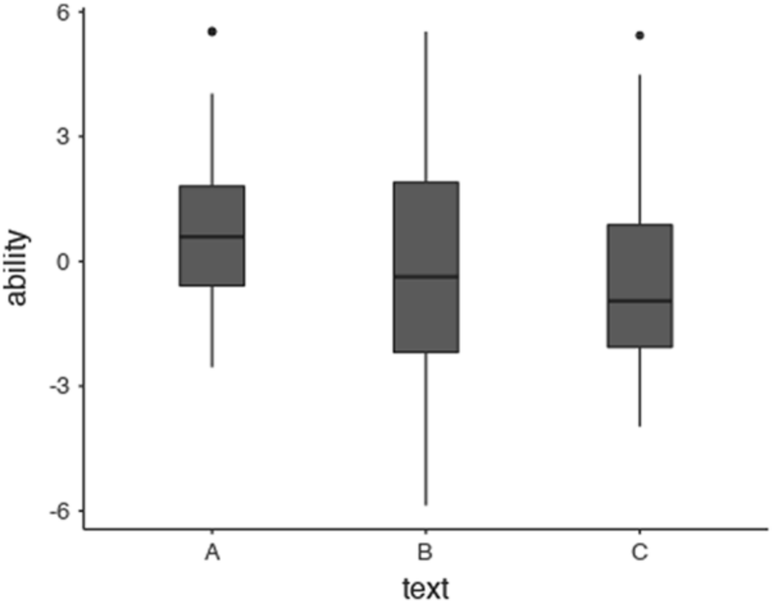


Fig. 2 Boxplot showing the scores for the three stories rated through comparative judgment. (ability = score assigned to each representation after the comparative judgment procedure)

Table 2 Descriptive statistics showing the spread in the rank order of the representations per picture story

	Min	Max	Mean	SD
Story A	−2.54	5.53	0.81	1.88
Story B	−5.87	5.52	−0.18	2.76
Story C	−3.97	5.43	−0.54	2.16

of Reference for Languages (Council of Europe, 2009). If the learners were unable to answer in English or did not write anything at all, their writing was scored as 'no output' which was considered a fifth level. The top and bottom level benchmark texts corresponded to representations that were ranked very high (5.52) or very low in the comparative judgment procedure (score of −3.8 and lower). For the bottom level there is no benchmark text as this level corresponds with texts written in Dutch or tasks where participants did not write anything at all. The benchmark text which corresponds with the A2-level received a score of 0.34 in the comparative judgment procedure, the A1-level benchmark text corresponds with a score of −0.23, and the benchmark text that was chosen for the below-A1-level corresponds with a score of. −1.16. The scores of the benchmark texts show that the rank order of the comparative judgment procedure was respected in text selection. Following the procedure Humphry and Heldsinger (2020) used for the assessment of L1 writing, performance descriptors were added to the benchmark texts. These descriptors are meant to give the characteristics of a text at a certain level and can help teachers when they are in doubt about which benchmark text is closest to their students' writing. The benchmark texts and descriptors together should give teachers the tools they need to assess similar writing tasks. The performance descriptors can also be used to give feedback to the students. The performance descriptors in this study were based on the CEFR descriptors for young learners aged 11–15 years (Goodier & Szabo, 2018). The benchmark texts and the descriptors for all levels can be found in Appendix B.

5.2 *Study 2*

5.2.1 Methodology

Participants

In this second study, 407 pupils from three schools participated. The study was conducted in schools that did not participate in study 1 (cf. 15.5.1). All participants were in the first year of secondary school. They were at the very start of formal L2 English education and had received 0 to 5 h of formal English classroom instruction. Each participant did the complete online skills test.

Instruments and Procedure

In this study, the reliability and efficiency of rating learners' writing with the benchmark texts were investigated using 407 learners' written texts. The writing task, a picture narration task based on picture story B, was given to the participants as the third task of our proficiency test measuring the four skills. Listening and reading skills were tested before the writing task, speaking was tested last. The students did the test on a desktop or a laptop, depending on what was available in their own school. They saw the visual as shown in Fig. 1. As the learners are at the start of L2 English lessons, instructions were given in English and Dutch. They were asked to type the story in a text box below the picture. Two raters scored the writing tasks using the benchmark texts with performance descriptors.To further investigate the efficiency of assessing written tasks with benchmark texts, we did an exercise with 98 teachers during a training session in which the tool was presented.

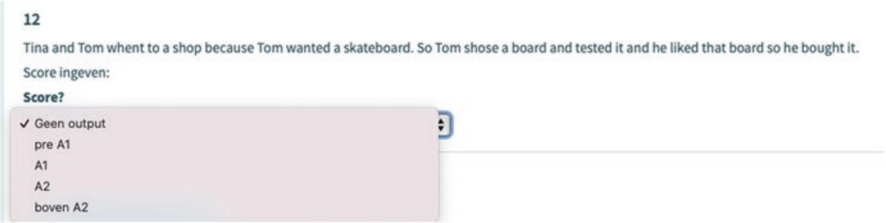
Analysis

We used descriptive statistics to report the spread in results and used weighted kappa to calculate interrater reliability. To report about the efficiency of scoring with the benchmark texts, we report the time raters needed to do the scoring activity.

5.2.2 Results

As mentioned above, two raters scored 407 written texts with the benchmark texts. The results showed that about 10% of the participants produced no output, 30% of the writing tasks received a pre-A1 score, 36% scored A1, 17% received an A2 score and 6% had a score higher than A2. Overall, 40% of the learners were not able to get the message across but 60% of the participants were already able to write a simple story at an A1 or A2 level. To investigate the reliability of the scores, 106 writing tasks were scored by both raters. The interrater reliability was high (weighted kappa = .90), indicating a very strong agreement between the raters' judgments. The two raters reported they scored about one task per minute, which showed this form of assessment to be very efficient.

To further check the reliability of using the benchmark texts with descriptors, 98 teachers, who took part in a session on how to use the online tool, did an activity where they rated 15 written texts with the benchmark texts. They were asked to first do this individually and then compare the results with a partner. The teachers were able to do the rating in 15 min. When comparing their ratings to those of another teacher, most ratings turned out to be the same and if they were different, they were



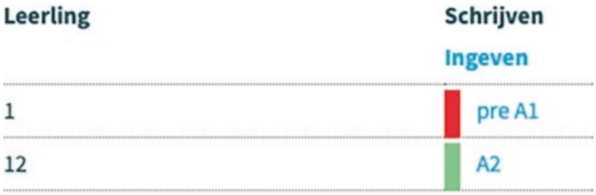
12

Tina and Tom whent to a shop because Tom wanted a skateboard. So Tom shose a board and tested it and he liked that board so he bought it.

Score ingeven:

Score?

Fig. 3 Screenshot from the teacher view in the tool showing the pupil's writing and a drop-down menu in which the teacher can add the score after comparison with the benchmark texts. (Dutch: Score ingeven = English: enter score, Dutch: Geen output = English: no output)



Leerling	Schrijven Ingeven
1	🟥 pre A1
12	🟩 A2

Fig. 4 Screenshot from the teacher view in the tool showing the pupils' writing scores. (Dutch: Leerling = English: pupil, Dutch: Schrijven = English: writing, Dutch: Ingeven = English: enter – this is where the teacher can click to see the pupils' writing and enter the score cf. Figure 3)

never more than one level higher or lower, which is in line with the reliability found above.

In the online tool which was made available for the teachers, the same procedure can be followed. The teacher can access the students' results via a results tab and can then access and score the writing with the help of the benchmark texts from the teacher's manual (cf. Figure 3). When all texts have been scored by the teachers, the teacher can consult the scores in an overview (cf. Figure 4). The score on the written tasks is not directly communicated to the pupils via the tool but via the teacher as was advised by the experts in the focus group on test development. A lower score means that pupils have less prior knowledge which is not necessarily a bad thing. However, it could be perceived as a failure by the learner. If teachers communicate the score to the pupils, they can better explain what the score means.

6 Discussion

In the study reported in this chapter, we investigated what a test meant to map pre-adolescent learners' L2 English proficiency can look like. From the results in the questionnaire it was clear that teachers wanted the test to look into young L2 English learners' proficiency in the four language skills (reading comprehension, listening comprehension, speaking and writing). This could be because the official curricula in Flanders stress the importance of language skills and focus on pupils' abilities to communicate. When asked about the test format, the teachers had a preference for an online tool. Furthermore, we opted for an online tool also because a website is easily accessible for all teachers.

The second part of this chapter focuses on using benchmark texts gathered via a two-stage approach as a method to rate L2 writing tasks designed for young learners in an online environment. The results show that benchmark texts which are selected via a two-stage approach, which was shown to be a reliable approach for L1 writing (De Smedt et al., 2020; Humphry & Heldsinger, 2020; McGrane et al., 2018), also leads to a reliable assessment of L2 picture narration tasks in a test designed for novice L2 English learners. One of the main advantages of this approach is that it is straightforward and easy to use for teachers. This is in line with observations in previous research (Lesterhuis et al., 2017; Humphry & Heldsinger, 2020).

We also aimed to investigate whether and how benchmark texts gathered via a two-stage approach can be integrated into an online tool that aims to assess L2 English writing skills. It was shown that the benchmark texts resulting from this approach can be integrated into an online tool that aims to map pupils' prior knowledge at the start of formal instruction. In the future, a similar approach could be followed for writing at a different level or with a different type of task and the use of benchmark texts could be integrated into other online tools in a similar manner.

The process leading to the choice of representative benchmark texts is quite time-consuming because a large group of raters and a lot of representations are necessary for the comparative judgment procedure to render highly reliable results. However, once this step has been taken, benchmark texts have been selected, and the descriptions of the different levels have been added, this method is very straightforward. The teachers are thus offered an efficient and reliable tool for rating their pupils' work. Benchmark texts are easy to use because teachers only need to compare their students' writing to the four available texts (with performance descriptors) or a fifth 'no output' option. They then decide which of the benchmark texts is of a similar quality to the students' writing. Once the teacher is familiar with these texts, the assessment can be done quickly and reliably. This approach makes it possible to integrate productive tasks (here: writing) in an online assessment tool that can be reliably assessed by the classroom teacher. This means that once the, albeit time-consuming, procedure of the two-stage approach has been completed, there is no need for extra raters or a centralized rating system to assess the writing tasks in this online tool.

Furthermore, the descriptors which are added to the benchmark texts can be used by the teachers either to give collective and individual feedback on their pupils' writing or in the design of their lessons. If, for example, the learners' writing tasks in a class group show large differences in prior knowledge, the teachers could decide to integrate these results in their lessons. They could offer materials to improve learners' writing (e.g. vocabulary necessary for the writing task or information on linking ideas in writing) which can appeal to all the learners in the group (e.g. through differentiated instruction).

Teachers or other stakeholders could also use this two-stage approach to select benchmark texts for other types of classroom assessment. They could either use one of the online tools which are currently available for comparative judgment and then follow the procedure described in this article for the selection of benchmark texts and performance descriptors.

If this approach is too time-consuming or expensive, they could also decide to look into a 'light version' of this two-stage approach. Teachers and teams of teachers could rank learners' writing tasks which they have rated in previous years in the first step and in the second step they could choose a number of tasks which they believe are representative as a benchmark for a certain level and describe why these tasks are considered representative (based on the objectives formulated by, for example, the curriculum or the CEFR). This method would give teachers in the same team the opportunity to all use the same benchmark texts with descriptors to assess their learners' writing similarly. Choosing a selection of representative benchmarks might add to the reliability of assessment in a team of teachers but the reliability of this 'light version' would have to be investigated in future studies. Furthermore, deciding on the performance descriptors might be an interesting exercise to do with a team as it might lead to a more deliberated assessment of students' writing.

In a follow-up study researchers could develop materials for teachers to tackle these differences in their L2 English classes but this was not within the scope of our project, which focused on the development of an online tool to assess learners' prior knowledge. Future studies could also investigate the differences between scores given using rubrics and scores given via the two-stage approach.

A limitation of this study is that the group concept mapping was done via an online questionnaire but the results from the online group concept mapping procedure were confirmed in the live focus group.

7 Conclusion

In this chapter we have shown how we decided on the content and form of an online tool to assess L2 English learners prior knowledge based on needs that were formulated by teachers. This resulted in a tool with tasks for all four language skills which can be completed in one 50-min lesson period.

One of the biggest challenges in the development of the tool was finding a reliable and efficient way to assess learners' written tasks. We decided to explore the possibility of using benchmark texts gathered in a two-stage approach. Overall, the two-stage approach combining comparative judgment and benchmark texts showed to be a good method to ensure reliable results when rating beginners' L2 narrative writing. This was shown by the interrater reliability in the second part of the study. Furthermore, the use of benchmark texts for assessment is straightforward and leaves little room for interpretation by individual teachers as there is one single holistic judgment based on comparison with a given set of texts (and descriptors). This proves to be a good method for assessing writing skills in an online tool, as, once the two-stage approach has been completed and the tool is online, the entire rating process can be done by the L2 English teacher. There is no need for external raters to score the writing tasks and the assessment still leads to reliable scores.

This study reported on an exploration into a holistic way to assess learners' writing and could be useful for teachers and other stakeholders who are looking for a practical, time-efficient, and reliable manner to rate their learners' writing. Further research with learners from different proficiency levels and different ages is warranted for the approach to be more widely used.

Notes The tool can be found here: https://www.starttoetsengels.be

Appendix

Appendix A: Teacher questionnaire

1. Highlight the correct answer:

You are a	man
	woman
	x.

2. Highlight the correct answer:

You are a	L2 English teacher in lower secondary education.
	L2 English teacher in higher secondary education.
	L2 English teacher in primary education.
	Teacher in primary education.
	Teacher in secondary education.
	Other: ____________________

3. Highlight the correct answer. How much teaching experience do you have?

   0–1 year
   1–3 years
   3–5 years
   5–10 years
   10–20 years
   20–30 years
   More than 30 years

4. How important do you consider the following statements? Give a score between 1 (completely unimportant) and 5 (very important) or answer not applicable.

   - This test should focus on realistic and real language.
   - This test should focus on academic language.
   - This test should focus on productive language.
   - This test should focus on receptive language.
   - This test should be linked to the CEFR.
   - This test should measure the four skills.
   - This test should measure listening skills.
   - This test should measure reading skills.
   - This test should measure speaking skills.
   - This test should measure writing skills.
   - This test should measure grammatical knowledge.
   - This test should measure lexical knowledge.
   - This test should measure student motivation.
   - This test should measure the pupils' attitude towards language(s) as a school subject.
   - This test should give feedback to the pupils.
   - This test should give teachers the opportunity to give feedback to the pupils.
   - This test should be done on paper.
   - This test should be done in a digital manner.
   - This test should be computer-adaptive.
   - This test starts with easy activities and gradually becomes harder in order to be able to find out the pupils' language level.

5. What is most important to you? Give a score from 1 to 5.

   The variation in skills which is measured (1) or the duration of the test (5).

6. How long can the test take?

   - 30 min
   - 50 min
   - Longer than 50 min

7. Do you have any other remarks or issues/worries you would like to see addressed?

Appendix B: Benchmark texts and descriptors for all levels

Benchmark text	Descriptors
Above A2 *Thomas went to the shop with his mom. After Looking at all the skateboards he asked Jon the shopkeeper, “can I try that one”. And he pointed at a beautiful black and green skateboard.* *“Of course” Jon said. And 10 seconds later Thomas was skating in the store.* *Wow i will buy it Thomas said, handing his money to Jon. How nice said his mom, go skate all te way home. Ha ha, said Thomas and so he did he skated all the way home.* *The End*	–The **message is clear.** –Overall high level output. –Grammar is mostly correct. Multiple tenses are used (‘went’, ‘will buy’). –Varied vocabulary (‘shopkeeper’, ‘handing money’, ‘all the way home’). –Linking words and conjunctions are used (more than only ‘and’) –The text is creative. The amount of text is higher than expected. The learner is not afraid to take risks. (‘handing his money to Jon’, ‘after looking at all the skateboards…’) –Mistakes result from taking risks (using language which is of a higher level, but maybe not yet completely mastered).
A2 *Tina and Tom whent to a shop because Tom wanted a skateboard. So Tom shose a board and tested it and he liked that board so he bought it.*	–The **message is clear.** –Overall the output is rather short. –Grammar: the writer tries to use different tenses and is often successful when doing this (e.g. ‘whent’, ‘shose’, ‘tested’). The grammar use does not lead to misunderstandings. –Vocabulary range is rather basic. Variation in choice of words is quite limited (e.g. repetition ‘board’) and the writer often chooses words which are related to Dutch (e.g. cognates). –Simple linking words and conjunctions are used in a correct manner. The writers goes beyond coordination (more than just ‘and’). –The text is not very creative. The text length is rather short.
A1 1. *A boy an a girl go’s to a skatboardshop* 2. *The boy asks if he kan trie one* 3. *He tries one* 4. *He likes it an then boat*	–The **message is clear.** –Overall the output is rather limited. –A lot of grammatical mistakes. The learner tries to use the correct form of tenses but often makes mistakes. (‘go’s’) –Vocabulary range is rather limited. Words which are known often are similar to the Dutch translation. (‘skateboard’, ‘shop’) –Spelling: more than a letter which is missing. Spelling is often based on pronunciation. –Use of linking words and conjunctions is limited. (not more than ‘and’ or ‘or’) –The text is not creative. The amount of text is limited.

(continued)

Below A1 1. *hmmm…* 2. *O hi.* 3. *It's really cool!* 4. *Perfect*	–The **message is not clear.** –The writer uses English but the meaning of the text is unclear without the visuals. Understanding comes from the interpretation of the reader rather than from the skills of the writer. –Grammar is mostly wrong or the amount of text is very minimal and it is hard to tell whether it is correct or not. –Vocabulary range is limited. Words show a clear link with Dutch words. ('cool', 'perfect') –Spelling: more than a letter which is missing. Spelling is often based on pronunciation. The amount of text is so minimal, it is hard to comment on the spelling. –Linking words and conjunctions are not or hardly used. –The text is not creative. The amount of text is limted.
No output *No text*	–No or hardly any English was used in the text. –Insufficient overall.

References

Arteveldehogeschool. (2021, October 5). *Starttoets Engels* https://www.starttoetsengels.be

Bouwer, R., Koster, M., & van den Bergh, H. (2016). *Benchmark rating procedure: Best of both worlds? Comparing procedures to rate text quality in a reliable and valid manner. In Bringing writing research into the classroom*. University of Utrecht.

Coenen, T., Coertjens, L., Vlerick, P., Lesterhuis, M., Mortier, A. V., Donche, V., Ballon, P., & De Maeyer, S. (2018). An information system design theory for the comparative judgement of competences. *European Journal of Information Systems, 27*(2), 248–261. https://doi.org/10.1080/0960085X.2018.1445461

Comproved. (2021, October 5). https://comproved.com

Council of Europe (Ed.). (2009). *Common European framework of reference for languages: Learning, teaching, assessment (10. printing)*. Cambridge University Press.

Crusan, D. (2013). Assessing writing. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 201–215). John Wiley & Sons, Inc.. https://doi.org/10.1002/9781118411360.wbcla067

De Bruyckere, P. (2017). *The ingredients for great teaching*. Sage.

De Smedt, F., Graham, S., & Van Keer, H. (2020). "It takes two": The added value of structured peer-assisted writing in explicit writing instruction. *Contemporary Educational Psychology, 60*, 101835. https://doi.org/10.1016/j.cedpsych.2019.101835

De Wilde, V., Brysbaert, M., & Eyckmans, J. (2020). Learning English through out-of-school exposure. Which levels of language proficiency are attained and which types of input are important? *Bilingualism: Language and Cognition, 23*(1), 171–185. https://doi.org/10.1017/S1366728918001062

Enever, J. (2011). *ELLiE – early language learning in Europe*. British Council.

Goodier, T., & Szabo, T. (2018). *Collated representative samples of descriptors of language competences developed for young learners*. Eurocentres.

Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In B. Kroll (Ed.), *Second language writing* (pp. 69–87). Cambridge University Press. https://doi.org/10.1017/CBO9781139524551.009

Harsch, C., & Martin, G. (2013). Comparing holistic and analytic scoring methods: Issues of validity and reliability. *Assessment in Education: Principles, Policy & Practice, 20*(3), 281–307. https://doi.org/10.1080/0969594X.2012.742422

Hattie, J., & Yates, G. C. (2013). *Visible learning and the science of how we learn*. Routledge.

Humphry, S., & Heldsinger, S. (2020). A two-stage method for obtaining reliable teacher assessments of writing. *Frontiers in Education, 5*, 6. https://doi.org/10.3389/feduc.2020.00006

Jones, N. (2016). 'No More Marking': An online tool for comparative judgement. *Research notes Cambridge English Language Assessment 63*, 12–15.

Lesterhuis, M., Verhavert, S., Coertjens, L., Donche, V., & De Maeyer, S. (2017). Comparative judgement as a promising alternative to score competences. In E. Cano & G. Ion (Eds.), *Innovative practices for higher education assessment and measurement*. IGI Global.

Lindgren, E., & Muñoz, C. (2013). The influence of exposure, parents, and linguistic distance on young European learners' foreign language comprehension. *International Journal of Multilingualism, 10*(1), 105–129. https://doi.org/10.1080/14790718.2012.679275

McGrane, J. A., Humphry, S. M., & Heldsinger, S. (2018). Applying a Thurstonian, two-stage method in the standardized assessment of writing. *Applied Measurement in Education, 31*(4), 297–311. https://doi.org/10.1080/08957347.2018.1495216

Muñoz, C., Cadierno, T., & Casas, I. (2018). Different starting points for English language learning: A comparative study of Danish and Spanish young learners: Different starting points. *Language Learning, 68*, 1076–1109. https://doi.org/10.1111/lang.12309

Pollitt, A., & Murray, N. L. (1996). What raters really pay attention to. *Studies in Language Testing, 3*, 74–91.

Puimège, E., & Peters, E. (2019). Learners' English vocabulary knowledge prior to formal instruction: The role of learner-related and word-related variables. *Language Learning, 69*(4), 943–977. https://doi.org/10.1111/lang.12364

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing, 22*(1), 1–30. https://doi.org/10.1191/0265532205lt295oa

Stoyanov, S., & Kirchner, P. (2004). Expert concept mapping method for defining the characteristics of adaptive E-learning: ALFANET project case. *Educational Technology Research and Development, 52*(2), 41–54. https://doi.org/10.1007/BF02504838

Thurstone, L. L. (1927). Psychophysical analysis. *American Journal of Psychology, 38*, 368–389.

Trochim, W. M. K. (1989a). An introduction to concept mapping for planning and evaluation. *Evaluation and Program Planning, 12*, 1–16. https://doi.org/10.1016/0149-7189(89)90016-5

Trochim, W. M. K. (1989b). Concept mapping: Soft science or hard art? *Evaluation and Program Planning, 12*, 87–110. https://doi.org/10.1016/0149-7189(89)90027-X

Verhavert, S., De Maeyer, S., Donche, V., & Coertjens, L. (2018). Scale separation reliability: What does it mean in the context of comparative judgment? *Applied Psychological Measurement, 42*(6), 428–445. https://doi.org/10.1177/0146621617748321