Digitally Assessing Text Comprehension in Grades 3-4: Test Development and Validation

Susanne Seifert and Lisa Paleczek
University of Graz, Austria
susanne.seifert@uni-graz.at
lisa.paleczek@uni-graz.at

Abstract: A prerequisite for child reading support at school is adequate assessment. Embedding (repeated) assessment into daily teaching routine is often challenging for teachers in terms of time and organization. The use of digital tools can help teachers in the assessment process (in preparation, evaluation, documentation, etc.). A digital assessment tool (Graz Reading Comprehension test: GraLeV), focusing on assessing reading comprehension skills in Grades 3 and 4 is currently being developed in Austria. This reading assessment covers reading comprehension at the word, sentence, and text level. Text level is assessed via two subtests (Subtest I: presentation of nonsense-stories and corresponding questions, and Subtest II: maze selection). The other levels consist of one subtest each. This paper focusses on the subtests at text level. More specifically, the paper reports the results of two studies. Study 1 describes the development phases and the first piloting of these two subtests (data collection: 10/2019-12/2019). Testing 273 students with preliminary versions of the subtests (Subtest I: 30 items, Subtest II: 60 items) produced information on (a) item difficulty, (b) item discriminatory power, and (c) time limits for future speed testing. Items not meeting the required quality criteria were excluded. The final version of Subtest I consists of 16 questions referring to eight different, short, nonsense-texts. Its testing time (without instructions) is three minutes. The final version of the Subtest II consists of 2 texts each with 15 maze selections (30 items) and testing time is 100 seconds. The internal consistency is found to be good for Subtest I (α=.87) and Subtest II (α=.78 to .80). Study 2 reports on testing for validity and retest-reliability (data collection: 09/2020-11/2020). Student scores in another reading comprehension test, together with teacher assessments of reading comprehension, were used to assess congruent validity. Divergent validity was assessed using teacher assessments of mathematical and socio-emotional skills. As expected, the correlations with the congruent measures were higher than those with the divergent measures. A subsample was tested twice with the GraLeV. Retest-reliability was acceptable for Subtest II. However, the scores obtained at time 2 were higher compared to those at time 1 in both subtests. This is probably the result of increased student familiarity with the digital device and the digital test environment at time 2. The results are discussed in the light of teachers' needs for standardized digital assessments in order to facilitate the tailoring of student reading support.

Keywords: reading skills; text comprehension; digital assessment; test development; quality criteria

1. Introduction

In today's world, reading skills are essential for accessing various forms of media and for the independent acquisition of knowledge. It is therefore no surprise that schools place such a heavy focus on reading ability. Once basic reading skills are acquired, the focus shifts towards gaining speed and improving comprehension so that knowledge may more easily be acquired across all subjects (Chall, 1983). Thus, text comprehension is a crucial ability. As text comprehension remains an essential skill throughout life, it is imperative that students with potential reading difficulties receive timely and adequate support. When assessing student skills, it is frequently the case that teachers use subjective methods and personal judgement rather than rely on standardized tests (Artelt and Gräsel, 2009). This makes it hard to avoid teacher bias, and explains why numerous studies have attempted to measure the accuracy of teachers' judgments (i.e., how well teacher judgement correlates with results obtained in standardized tests). Südkamp, Kaiser and Möller (2012) found in their meta-analysis (including 75 studies) an overall correlation of 0.63 between teacher judgment and student achievement. With respect to reading skills, Paleczek, Seifert and Gasteiger-Klicpera (2017) have shown that certain aspects such as class size or student ability and background may have a negative influence on the accuracy of teachers' judgments. Using standardized tests regularly could therefore enhance judgment accuracy. Thus, in this context, the use of diagnostic procedures to regularly assess reading skills in a simple, straightforward manner during class time plays a crucial role. Such procedures not only aid teachers in identifying students with reading deficiencies, they also ease the regular checking of support measure success.

Some digital tools for assessing German reading skills are only available for purchase (e.g., ELFE II: Lenhard, Lenhard and Schneider, 2020; ProDi-L: Richter et al., 2017; VSL: Walter, 2013), and others (LEVUMI: Mühling and Gebhardt, 2021) are available for free. However, these open access tools are only rarely available in a standardized form. In this paper, we present two subtests from a new standardized, open access diagnostic tool, the Graz reading comprehension test (Grazer Leseverständnistest: GraLeV; Paleczek et al., in prep.). The GraLeV

Reference this paper: Seifert, S., and Paleczek, L., 2021. Digitally Assessing Text Comprehension in Grades 3-4: Test Development and Validation. *The Electronic Journal of e-Learning,* 19(5), pp. 336-348, available online at www.ejel.org

focuses on assessing reading skills in Grades 3 and 4. Since it is digital (tablet, notebook, PC), it provides an efficient means of assessing reading skills (with a pure test time of about 10 minutes). Four different subtests provide differentiated information on student reading skills at the word, sentence and text level. So far, the test has been conceptualized, developed, and digitalized, and has also gone through the initial piloting procedures necessary for assessing test quality criteria and usability.

The present paper focusses on the two subtests at the text level (Subtest I and II). First, the different methods for assessing reading comprehension (at text level) are introduced. Second, the advantages and drawbacks of digital assessments are described. We then report on the conception and construction of the digital reading test and follow this with details of study results. Study 1 reports on a piloting study in which we gained insight into (a) item difficulty, (b) item discriminatory power, (c) internal item consistency, and (d) resulting time limits. Study 2 describes subtest validity and reliability.

2. Assessing reading comprehension in primary school

Reading skills generally pertain to a person's ability to successfully cope with certain types of textual or reading-related demands (Artelt et al., 2007, p. 11). This involves various sub-processes of reading, some of which are automatic, and some of which are deliberately controlled (Graesser, Singer and Trabasso, 1994). For most children, the systematic acquisition of reading skills begins with school entry. Initially, basic reading skills (including decoding, reading comprehension at word and sentence level) are acquired. Then, increasingly complex reading and comprehension processes are trained (at the latest from Grade 4 onwards) while reading fluency and speed increase (Klicpera et al., 2017). In particular, the ability to understand texts when reading them is crucial to enabling learning from texts (Schnotz, 1994), a skill which is expected of students at the end of primary school. When reading texts comprehensively, it is necessary to interpret units of meaning across sentences and to establish local and global coherence (Richter and Christmann, 2009).

In primary school in German speaking countries, different test procedures are used to assess reading comprehension skills (for an overview: see Lenhard, 2013; Paleczek and Seifert, 2019). Teachers can use short diagnostic procedures to rapidly gain a rough overview of their class's reading status and to identify students with reading deficiencies (e.g., SLS 2-9: Mayringer and Wimmer, 2014). Some procedures are designed as formative assessment (e.g., VSL: Walter, 2013). Teachers may also use more comprehensive diagnostic tools to test various sub-processes of reading (and especially reading comprehension) (e.g., ELFE II: Lenhard, Lenhard and Schneider, 2020; HAMLET 3-4: Lehmann, Peek and Poerschke, 2006). Word, sentence and text level represent sub-processes and are often dealt with in such tests.

In order to monitor learning progress, teachers need simple procedures capable of measuring reading skills accurately (Guthrie et al., 1974). One way of doing this is to let students answer *questions about a text*. The difficulty of this task, however, varies according to question type (e.g. direct information extraction vs. inference extraction) and answer mode (yes-no questions vs. open questions) (Guthrie et al., 1974). In these procedures in German speaking countries, the answer mode is often a multiple-choice format (e.g., ELFE II: Lenhard, Lenhard and Schneider, 2020). Unfortunately, as students are often already familiar with entities addressed in the texts and questions (animals, plants etc.), their answers also reflect their background knowledge and may not be a pure measure of reading comprehension.

Another possibility of measuring text comprehension is the *maze procedure*, which is often mentioned and used in progress diagnostics of reading comprehension. The maze procedure is a universal means of screening in order to identify weak readers. This procedure has already been extensively tested, especially in English-speaking countries (for an overview: see Wayman et al., 2007). In such a procedure, students are asked to quietly read a section of a text within a certain time limit. In the text, at every seventh word, a target word must be identified out of three presented words (2 distractors, 1 target word). By setting a time limit, the maze procedure is a good way of measuring reading fluency or speed (Muijselaar et al., 2017). By increasing the level of distractor difficulty, the maze procedure can also capture reading comprehension (Conoyer et al., 2017). There are, however, limitations when it comes to measuring reading comprehension. Some researchers argue that the maze procedure assesses comprehension at the sentence rather than the text level (Gellert and Elbro, 2012). In German-language reading tests, the maze procedure is used, for example, in the VSL (Walter, 2013) for Grades 1 to 6.

3. Digital assessments in primary school

In order to provide each student with tailored support, and to regularly check on their learning progress in reading instruction, it is necessary to assess reading (sub-) abilities and individual learning. As this is all rather time consuming, teachers find it difficult to embed such assessment into their daily teaching routine. The use of digital testing procedures can thus support teachers in the diagnostic process, especially with respect to the preparation, implementation, evaluation, and documentation of assessments (Cheung and Slavin, 2012; Neumann et al., 2019). Students can then receive more rapid feedback, teachers save time in follow-up (Ehlers et al., 2013) and spend time on instruction rather than on testing and documentation (Gebhardt, Diehl and Mühling, 2016). Clearly, all such tools need to be easy to use (i.e., to exhibit high usability, as understood by Nielsen (1993)), and schools need to be adequately equipped (e.g., with tablets, reliable class internet, etc.), before such benefits can be realized.

To an increasing extent, standardized reading tests now offer a digital version in addition to the classic paper-and-pencil version. These digital versions facilitate or automate implementation and evaluation. In the area of German reading tests, for example, the ELFE II (Lenhard, Lenhard and Schneider, 2020) is one such hybrid procedure. There are also some tests that can only be performed digitally, such as the ProDi-L (Richter et al., 2017). While studies on these tests examined the scientific reliability and internal validity of the test instruments or looked at the mode effect (Lenhard, Schroeders and Lenhard, 2017; Richter et al., 2017), the question of usability (Nielson, 1993) was not dealt with. A well-designed interface enabling intuitive use is a clear prerequisite for any digital tools. Although usability studies, e.g., the aspect of testing for subjective user satisfaction, have only rarely been part of the educators' research tradition, they have recently become more common in the evaluation of German digital assessment tools. For example, studies have generally shown that digital tests have been very well accepted by primary school teachers (Förster and Souvignier, 2014). For a digital maths tool, Blumenthal and Blumenthal (2020) analysed German fourth graders' opinions on print and digital modes of testing. They state that students found digital testing to be more motivating than testing in print form, and that students perceived digital tests as being faster and easier (although this perception was not corroborated by the test data). Similarly, preferences for digital mode were revealed in assessing vocabulary skills in kindergarten children (Paleczek, Seifert and Schöfl, 2021).

While digital test possibilities appear promising, several variables still require particular attention. For example, not all teachers are used to using digital devices in their lessons (Brandhofer, 2015), and many teachers do not feel competent enough to guide students in their use (Schaumburg, 2015). Student digital competence also needs to be considered. Elementary school students, in particular, often do not have the skills needed to deal with a digital device on their own (Medienpädagogischer Forschungsverband Südwest, 2018). Thus, the evaluation of usability, although long acknowledged as essential with respect to human-computer interaction (Nielsen, 1993), still needs to be specifically addressed in the context of educational assessments.

4. Conceptualising the digital reading comprehension test at text level

Our goal was to design a comprehensive and cost-efficient digital test for assessing reading comprehension, one capable of aiding teacher decision-making and fostering individual student support. To carry this out, we planned subtests for assessing the three different levels of reading comprehension: word, sentence and text level. This paper focusses on text level. The conception and construction of the test and its items took place in summer 2019. The test was installed on a current HTML 5 platform. The exercise contents are stored in xml files which are read and interpreted by the Exercise Viewer using Angular 2 / Typescript. The subtests consist of exercise packages that can be given to students individually or in groups. The test was provided on a self-developed LMS.

For text-level, we decided to develop two different subtest formats, both of which are commonly used to assess reading comprehension on text-level: For Subtest I, we constructed a test that assesses whether students can extract information from short texts. Subtest II uses a maze procedure. The two approaches are described in detail below.

4.1 Subtest I: Questions regarding short nonsense-stories

Subtest I consists of various short texts, each followed by two multiple-choice questions. The texts present information on nonsense-things, creatures (both nouns) or actions (verbs). We used nonsense-content to obviate student use of background knowledge when answering questions.

The texts were constructed either as easy-to-read texts, containing two to three sentences (n=9), or as more demanding texts (e.g., including more sub-clauses, longer sentences) containing four to five sentences (n=7).

In analysing text readability for Grade 3 and 4 students, we applied the readability formula employed in the Regensburg Index (RIX: Wild and Pissarek, 2019). This incorporates characteristic values for readability (e.g., multi-syllabic words, number of sentences) and includes difficulty parameters (e.g., passive forms, sentence complexity). This readability formula has been tested for German texts and provides information on the suitability for certain grades. To determine text characteristics (see Table 1), we used the Regensburg analysis tool for texts (Ratte: Wild and Pissarek, n.y.).

Table 1: Subtest I before first piloting: Characteristics and RIX of texts

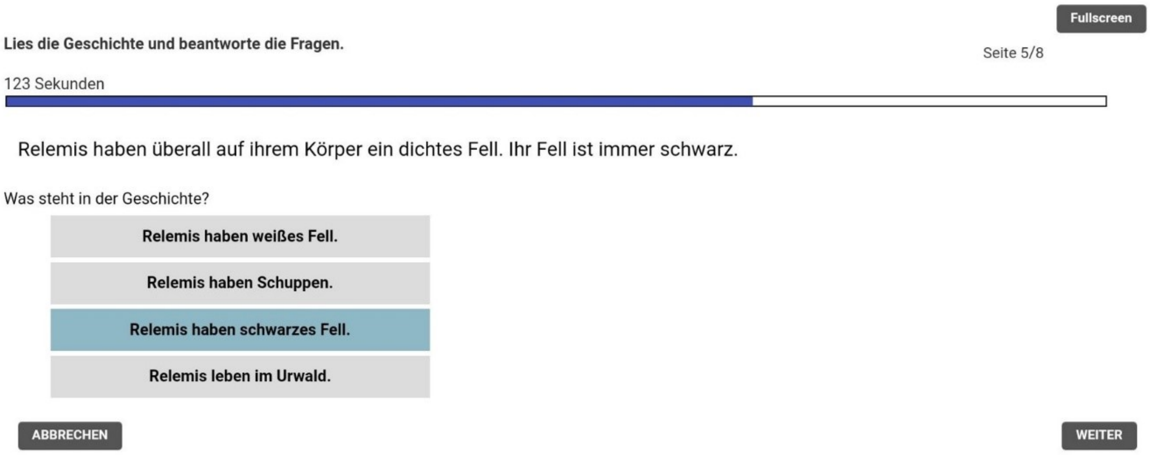
Text Type	**Text Number**	**Nonsense-word**	**Word Type**	**Word-count**	**Sentence-count**	**RIX**
short	1	Tinatos Kanat	creatures thing	17	3	2.13
	2	krolken	action	20	2	2.93
	3	Stasmir	thing	20	2	2.94
	4	minnern	action	16	3	2.09
	5	Delliwam	thing	26	3	2.74
	6	Relemis	creatures	14	2	2.44
	7	Rafiza	thing	12	2	2.22
	8	Basati	thing	18	3	2.22
	9	branteln	action	13	3	1.86
M (SD)				**17.33 (4.33)**	**2.56 (0.53)**	**2.40 (0.39)**
long	10	Sinalas	creatures	41	4	3.98
	11	Fenati	thing	41	3	4.47
	12	Zünglis	creatures	44	4	4.1
	13	Makentas	thing	38	5	3.54
	14	Wanila	creature	42	4	4.02
	15	Tentaris	creatures	52	5	4.41
	16	frijaben Frijabis	action thing	51	4	4.34
M (SD)				**44.14 (5.34)**	**4.14 (0.69)**	**4.12 (0.32)**

The multiple-choice questions corresponding to the texts covered the two important comprehension processes tested in international large-scale studies (e.g., end of Grade 4 in PIRLS: Widauer and Wallner-Paschon, 2017) and in well-established reading tests (e.g., ELFE II: Lenhard, Lenhard and Schneider, 2020). The first process requires retrieving explicitly stated information. The second process requires making straightforward inferences. It thus becomes possible to assess a child's basic level of text comprehension using just two questions for every text.

For the two questions offered, there was one correct answer (target) and three distractors. The distractors were constructed to be at least theoretically possible, thus encouraging students to concentrate on reading and understanding the text. To familiarize students with this type of text (nonsense-texts) and with the task format (two questions on each text), one text (Text 1) was used as an example. Teachers solve the sample questions together with their students to check student understanding of instructions.

In October 2019, we individually offered the texts with their corresponding questions (30 items) to six children (Grade 3: n=3, 2 boys, 1 average and 1 weak reader; 1 girl strong reader; Grade 4: n=1 girl strong reader, Grade 5: n=2 boys, very weak readers). We used observation and the think-aloud method while children conducted the subtests, as well as brief interviews after they finished the test, to obtain information on the solvability of the items and the tool's usability. Afterwards, three texts, questions and/or the answers to the corresponding questions were slightly modified to remove unnecessary ambiguity from the target items. In addition, the digital presentation of text and questions was changed. Initially, the text and both multiple-choice questions were presented on the same screen. This led to small letters and thus resulted in reduced readability, as was revealed by observation and reported by the children. After receiving children's feedback, we decided to present the text and the first question on one screen. On the subsequent screen, we presented the text again, this time with the second question. This also meant that each screen had to load, and thus increased dependence on the internet connection. Furthermore, test instructions were also adapted accordingly. Figure 1 shows the presentation mode of an item in Subtest I after adaptations.

Following these initial adaptations, a pilot test was carried out with a whole Grade 3 classroom (11/2019). A period of observation and classroom discussion after the test revealed that there was no further need for technical or item adjustment (in order to enhance usability).



Lies die Geschichte und beantworte die Fragen.

Relemis haben überall auf ihrem Körper ein dichtes Fell. Ihr Fell ist immer schwarz.

Was steht in der Geschichte?

- Relemis haben weißes Fell.
- Relemis haben Schuppen.
- Relemis haben schwarzes Fell.
- Relemis leben im Urwald.

Figure 1: Presentation mode of Subtest I (Short Text 6 with the first question, solved)

4.2 Subtest II: maze selection

For Subtest II, four different factual texts were selected from materials designed to foster Grade 3 students' reading skills (developed in the project LARS from 2012 to 2014, for more information see https://differenzierter-leseunterricht.uni-graz.at/). The topics were deliberately chosen from these materials to cover age-specific interests and to be gender neutral.

After selecting the four factual texts, the text length and the presumed time limit for the test were determined. To ensure that the whole text could be presented on a tablet screen in an appropriate font size without scrolling, we set text length to about 100 words. According to the norms of the SLRT II (Moll and Landerl, 2010), the upper quartile of Austrian Grade 4 students can read approximately 100 words per minute. Studies on maze procedures, however, have shown that higher test reliability goes hand in hand with increased time (Conoyer et al., 2017; Espin et al., 2010) and they suggest a time limit of three minutes (Conoyer et al., 2017). Combining two texts in consecutive screens would have led to students having to click on a "next-button" on the first screen. For us, this option was unsatisfactory as we did not want differences in student scores arising as a result of their ability to click a "next-button". We therefore decided to use texts of about 100 words at different levels of difficulty. In piloting these texts, we wanted to find out whether easy texts, more demanding texts or both would lead to satisfying test reliabilities. Hence, the four selected texts were slightly rewritten and shortened to offer two relatively easy texts (at grade level) and two relatively demanding texts. Two different formulas were applied in order to define readability level. One of them was the gSmog (Simple measure of Gobbledygook - German; Bamberger and Vanecek, 1984), which measures the number of multi-syllable words (more than three syllables) in relation to the number of sentences. The other formula was the RIX (Wild and Pissarek, 2019). Both readability formulas were tested for the German language and they allowed for selection of a suitable text for the grade in question. The tool Ratte (Wild and Pissarek, n.d.) was used to determine text characteristics. Based on these indices, two easier (Texts 1 and 2) and two more demanding texts (Texts 3 and 4) were then designed (see Table 2). Finally, we added questions stating the topic of the text as headings. The use of such headings was intended to arouse student interest and motivate them to read the respective text.

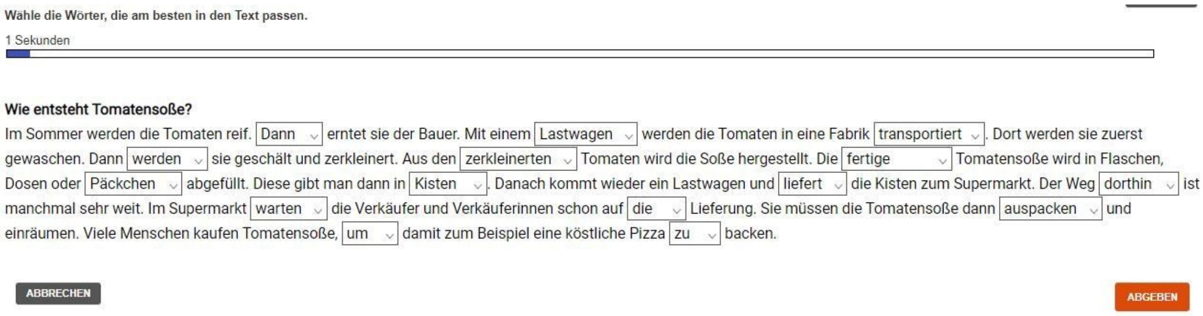
Table 2: Information on the four texts (including text characteristics, gSmog and RIX) in Subtest II

Text Number	Title	Word-count (without distractors)	Sentence-count	gSmog	RIX
1	What can we discover in nature?	104	12	3.48	4.72
2	What do you know about farm animals?	107	15	3.83	3.87
3	How is tomato sauce made?	106	13	5.29	5.29
4	Where do we get our food from?	104	15	6.37	4.45
M* *(SD)		**105.25 (1.5)**	**13.75 (1.5)**	**4.74 (1.34)**	**4.82 (0.43)**

In accordance with the maze procedure used in English-speaking countries (Fuchs and Fuchs, 1992), every seventh word (not counting the heading) was replaced by a drop-down option that contained the target item (fitting the text) and two distractors. Each text contained 15 drop-down options. As is typical for this task format (see Walter, 2013), we employed one distractor resembling the target word grapheme-phonemically and one distractor resembling the target word in a semantic-syntactical way. We ensured that the distractors (at least as far as the word types nouns, verbs and adjectives are concerned) were (a) grammatically correct, (b) syntactically possible, and (c) related to the context (Ketterlin-Geller et al., 2006). This served to raise the level of distractor difficulty. It has been found that increased distractor difficulty goes hand in hand with higher construct validity in assessing reading comprehension (Conoyer et al., 2017).

To familiarize the students with the task format, we prepared a text consisting of two sentences on "Why is the Earth called the blue planet?”. In this short text, three drop-down options were presented.

Finally, in October 2019, all of the four texts (15 items each, 60 items in sum) were tested in individual settings with six children using observations, the think-aloud method and interviews as described above. As subsequent review of the processing revealed uncertainties concerning five of the items, we adapted the distractors. The children had no difficulty with the test format, nor with its digital presentation, thus usability was given. Again, pilot testing with a whole Grade 3 classroom in November 2019 revealed no further need for technical or item adaptation (in order to enhance usability) for this subtest. Figure 2 shows Subtest II.



Wähle die Wörter, die am besten in den Text passen.

1 Sekunden

Wie entsteht Tomatensoße?

Im Sommer werden die Tomaten reif. Dann erntet sie der Bauer. Mit einem Lastwagen werden die Tomaten in eine Fabrik transportiert. Dort werden sie zuerst gewaschen. Dann werden sie geschält und zerkleinert. Aus den zerkleinerten Tomaten wird die Soße hergestellt. Die fertige Tomatensoße wird in Flaschen, Dosen oder Päckchen abgefüllt. Diese gibt man dann in Kisten. Danach kommt wieder ein Lastwagen und liefert die Kisten zum Supermarkt. Der Weg dorthin ist manchmal sehr weit. Im Supermarkt warten die Verkäufer und Verkäuferinnen schon auf die Lieferung. Sie müssen die Tomatensoße dann auspacken und einräumen. Viele Menschen kaufen Tomatensoße, um damit zum Beispiel eine köstliche Pizza zu backen.

Figure 2: Presentation mode of Subtest II (Text 2, with solutions)

5. Aims and research questions

After undertaking revision and finalizing the construction, we began piloting on the preliminary versions of the subtests to gain insight into the tests’ characteristics (Study 1). More specifically, by using the tests as power tests (excluding a time limit), we gained insight into each single item’s characteristics. Further adaptations were then made based on these insights: Items were deleted/ adapted, a time limit was set, instructions were adapted, and the test’s final length was decided upon. Study 2 was used to test the validity and reliability (the quality criteria) of the final test versions.

6. Study 1: Item analyses and internal consistency

6.1 Procedure

In their final versions, both subtests of the GraLeV are designed to be speed tests. To gain information on item characteristics and internal test consistency, however, it is necessary to conduct the tests as power tests (with no time limit). This enables the students to work on every item of both subtests. To avoid frustration among very weak readers, the test was abandoned once 80% of the students in class had finalized the subtest in question. The power test procedure provided enough information to analyse item difficulty and discriminatory power). The items' difficulty was calculated in Excel as corrected item difficulty, taking account of the number of distractors (Eid and Schmidt, 2014). Item difficulty was intended to be comparably low (with high coefficients of at least 0.6) since the final tests would be conducted as speed tests. The items' discriminatory power was determined by using the reliability analysis in SPSS, which provided a measure of internal test consistency using Cronbach's alpha. To ensure test reliability, an item's discriminatory power needs to be as high as possible, and never below 0.3. A distractor analysis was also conducted to identify any problematic distractors. All items that did not meet the quality criteria were excluded from the original final item set and the internal consistency of the resulting item set (as a measure of reliability) was then calculated. According to Bühner (2011), reliability values of 0.7 are considered as acceptable, and values above 0.8 are considered good.

The time students needed to work on each item was also recorded. This information helped in determining time limits for the speed tests.

Data collection took place in Grade 3 and 4 classrooms from October to December 2019. For each classroom, there were two different testing days (each day spending about one lesson in the classroom) as the preliminary power test version of the GraLeV consisted of a relatively large number of items. On day 1, students worked on the Subtest Word and the Subtest Text II. On day 2, students worked on the Subtest Sentence and the Subtest Text I. Project members (university staff and project interns) conducted the test. They were trained to guide the students through the test procedure. As internet service in Austria is inadequate in many classrooms, in a lot of cases mobile phone internet hotspots had to be used.

Teachers provided relevant information concerning student background.

6.2 Instruments

6.2.1 GraLeV

The GraLeV is a digital test constructed to assess reading comprehension at the word, sentence and text level. Text level is assessed via two subtests (as described above). The Subtest Word consisted in this version of two sample items and 38 test items. Each item consists of three pictures and six words. Students needed to pick the three words that fit the pictures and put them via drag and drop under the pictures (in a box). The Subtest Sentence consisted of 22 items. Each item presented a picture and four sentences below the picture. Students needed to pick the sentence that fit the picture (target sentence). The other three sentences were distractors.

6.2.2 Teachers' questionnaire

The teachers' questionnaire consisted of six questions concerning student background: month and year of birth, first language, special educational needs (SEN; if so: which, and taught by which curriculum) or extraordinary status (this is used in Austria for children who are second language (L2) learners with a very low ability to understand the language of instruction). Teachers filled out the questionnaires in between the testing days.

6.3 Sample

A total of 273 students took part in the study (with their parents' consent), 117 being Grade 3 students. When teachers stated that the student spoke at least one language different from German at home, the student was defined as a L2 learner. Table 3 provides details on the sample.

Table 3: Sample descriptives of Study 1

Grade	N	% female	% L2 learners	% SEN
3	117	49.6	9.4	0.9
4	156	48.7	30.2	2.6
total	273	49.1	21.2	1.8

6.4 Results

In Subtest I, between 235 and 241 students worked on the 30 items (15 stories with two questions each). After analysing items, we identified eight items with a difficulty below 0.6. Additionally, the distractors of three of these items were chosen by more than 20% of the students. Furthermore, one of these item's discriminatory power was below 0.2. These items were not included in the final item set. A text was only retained in the final item set when both corresponding questions met the quality criteria. Thus, seven texts were excluded, resulting in a final set of eight texts (six short texts, two long texts, and each with two questions: 16 items). In terms of Cronbach's alpha (α=0.87), the internal consistency of this final item set was good. Items, were then ordered according to their corrected item difficulty, beginning with the easiest item.

The time students needed to answer both questions for each text was recorded. Based on this, we calculated the average time needed for working on the 16 items of the final item set: 333 seconds (SD=118). The minimum time recorded for working on this item set (72 seconds) was not deemed reliable. We assumed that students had only clicked through answers without having read the text (e.g., needing only 2 seconds for solving both questions referring to Text 12). We thus drew on the time needed by the 25% fastest readers. These needed 258 seconds. Maximum times were also considered to define how long slow readers took for one text with two questions. We wanted students who are weak readers to be able to solve at least one or two questions. This also enables us to differentiate within the group of weak readers. Thus, the maximum time needed for the easiest text in the final item set (short text 6) was analysed. The slowest reader of the sample solved this item in 116 seconds. Based on these findings, we decided to set the time limit to 3 minutes. This limit would prevent the fastest readers from finishing all items, and additionally enable weaker readers to solve at least one item.

In Subtest II, 246 to 251 students worked on the four texts and, hence, on the 60 items. An analysis of the items revealed that Text 1 and Text 4 contained items with corrected item difficulty below 0.6 (two items), or discriminatory power below 0.2 (four items). Text 2 (easy) and Text 3 (more demanding) showed good item characteristics and were included in the final test version. The internal consistencies of Text 2 (Cronbach's α=0.80) and 3 (Cronbach's α=0.78) were satisfactory. Except for one item in Text 3, all items had acceptable values. Separate consideration of the item characteristics in Grade 3 and Grade 4 revealed, however, that in the sample of Grade 3 students, three further items in Text 2 showed discriminatory power below 0.2. Thus, to improve item characteristics in Text 2, one sentence was slightly modified, and three distractors were changed. In one item in Text 3, one distractor was changed.

Looking at the solution times needed by the 172 students revealed that about 167 seconds (*SD*=63.59) were required for Text 2, and 174 seconds (*SD*=62.22) for Text 3. The fastest readers needed 48 seconds for Text 2, and 66 seconds for Text 3. Again, we drew on the time needed by the 25% fastest readers, i.e. about 122 seconds and 130 seconds for solving Text 2 and Text 3, respectively. To ensure that slow readers also had time to work on at least some items, the maximum time recorded for the texts was divided by the 15 items per text. For Text 2, the slowest reader needed about 30 seconds for solving one item. For Text 3, the slowest reader took 29 seconds. Based on these findings, we decided to set the time limit at 100 seconds for reading both texts consecutively. This enables very slow readers to solve at least some items and prevents fast readers finishing both texts before the limit expires.

7. Study 2: Validity and reliability

7.1 Procedure

The final versions of both subtests were performed with students to gain information about the tests' (a) validity (convergent and divergent) and (b) reliability (retest-reliability). Convergent validity measures contained teacher assessment (TA) of student reading skills and student performance on the reading comprehension test ELFE II (Lenhard, Lenhard and Schneider, 2020). According to Bühner (2011), the correlation coefficient of the convergent validity needs to be above 0.5 in order to conclude that the tests assess the same ability.

Divergent validity was calculated based on the TA of students' mathematical and social-emotional skills.

Data collection took place in Grade 3 and 4 classrooms from September to November 2020. Again, we went to the classrooms twice to avoid overwhelming students (Day 1: GraLeV and ELFE II text; Day 2: GraLeV and ELFE II word and sentence). Project members conducted the tests. To be independent of school internet, we brought portable routers to the classrooms. Although this worked better than the hotspots in Study 1, using internet with the whole class still proved challenging. Unfortunately, COVID-19 measures prevented us from going to some classrooms twice. Some results on ELFE II word and sentence subtests are therefore missing. In the subsample tested on both days, we were able to gain knowledge of the GraLeV's retest-reliability. Reliability coefficients above 0.7 are considered as acceptable (Bühner, 2011).

7.2 Instruments

7.2.1 GraLeV

The GraLeV was used in its final version (after Study 1) as a speed test to digitally assess reading abilities on the word, sentence and text level. Text level is assessed via two subtests (as described for Study 1). The Subtest Word consisted of two sample exercises and 12 test exercises (time limit: three minutes). The Subtest Sentence consisted of 22 items (time limit: three minutes).

7.2.2 Teachers' questionnaire

The teachers' questionnaire consisted of the same six questions as in Study 1. Additionally, teachers had to assess students' reading skills (word, sentence and text), their mathematical skills (numerical understanding and spatial-visual skills) and their socio-emotional skills on a Likert-scale, with responses ranging from 1 (weak) to 7 (strong).

Teachers filled out the questionnaires while we were testing the students.

7.2.3 ELFE II

This test was used to measure reading ability at word (75 items), sentence (26 items) and text level (36 items). Reliability values are reported by Lenhard, Lenhard and Schneider (2020) to be r_{tt}=.93 (for retest) and *r*=.96 (for odd-even-split-half-reliability). We used the paper-pencil version of the test. The subtests had time limits (word: 3 minutes, sentence: 3 minutes, text: 7 minutes). On word level, students needed to choose one of four words that best fit the picture presented. At sentence level, students were presented with a sentence where one word had to be chosen out of five to fit the sentence. At text level, students read short texts and had to answer one to three questions on the text.

7.3 Sample

A total of 534 students took part in the study (with their parents' consent), 333 being Grade 3 students. When teachers stated that the student spoke at least one language other than German at home, the student was defined as L2 learner.

The amounts of test data differ due to the two measuring times used. We gathered information on the two GraLeV-subtest scores (time 1) from 447 (Subtest I) and 451 students (Subtest II). Data on scores in the ELFE II subtests is available for 357 (word level), 364 (sentence level) and 386 students (text level). Teachers' assessment on reading, mathematical and socio-emotional skills was available for 458 students. For a subsample of 169 students, retest data are also available (time 2). Tables 3 and 4 provide further details.

Table 3: Sample descriptives of Study 2

Grade	N	Age *M (SD)*	% female	% L2 learners	% SEN
3	333	8.78 (0.47)	43.0	33.5	0.7
4	201	9.82 (0.46)	54.8	43.4	1.3
total	534	9.14 (0.68)	47.1	37.0	1.0

Note: The information on age, gender, and language spoken at home was only provided for 337, 352, and 356 students, respectively. Percentage calculations are based solely on the data collected. The missing student data was ignored.

Table 4: Subsample retest descriptives

Grade	N	Age *M (SD)*	% female	% L2 learners	% SEN
3	112	8.72 (0.43)	48.2	25.9	0
4	57	9.84 (0.50)	56.1	29.8	1.8
total	169	9.10 (0.70)	50.9	27.2	0.6

7.3 Results

7.3.1 Validity

Table 5 shows the correlations of the two GraLeV-subtests on text comprehension with the ELFE II reading comprehension test and the TA of reading comprehension (convergent validity), as well as the correlations with the TA of mathematical and socio-emotional skills (divergent validity). As expected, the reading comprehension score obtained in the GraLeV's text comprehension subtests correlated highest with the ELFE II reading comprehension scores (.57 to .75). Likewise, the GraLeV scores correlated highly with TA of text comprehension (.40 to .53). Even though there are significant correlations of the GraLeV subtests with TA of mathematical skills (.26 to .40), these are mostly significantly lower than the correlations found for convergent validity (e.g., Subtest I correlates significantly higher with TA of text comprehension than with TA of spatial-visual skills: z = 1.74, p < .05). Only in Grade 4, there is no significant difference between the correlation of Subtest I with TA of text comprehension (convergent validity) and TA of spatial-visual skills (divergent validity): z = 1.36, p = .09). In general, the subtests' scores correlated lowest with TA of socio-emotional skills (.08 to .40, see Table 5).

Table 5: Validity

	ELFE II reading comprehension test			**TA of reading comprehension**			**TA of mathematical skills**		**TA of socio-emotional skills**
Grade 3	**word**	**sentence**	**text**	**word**	**sentence**	**text**	**Numerical understanding**	**Spatial-visual skills**	
Subtest I	.57**	.67**	.61**	.33**	.34**	.40**	.28**	.30**	.08
Subtest II	.63**	.74**	.70**	.38**	.41**	.46**	.26**	.27**	.08
Grade 4									
Subtest I	.61**	.65**	.67**	.43**	.46**	.47**	.34**	.38**	.24**
Subtest II	.69**	.74**	.67**	.50**	.52**	.53**	.33**	.40**	.40**
Total									
Subtest I	.63**	.70**	.69**	.33**	.35**	.40**	.31**	.33**	.16**
Subtest II	.69**	.76**	.75**	.39**	.41**	.46**	.29**	.32**	.21**

Note: ** p<.01

TAs were highly correlated with each other. As expected, TAs of different levels of reading comprehension correlated highly with each other (.87 to .96), as did the TAs of the two mathematical skills (.77). Interestingly, there were also high inter-correlations between the TA of reading skills and the TA of mathematical skills (.51 to .61) and between those of reading skills and socio-emotional skills (.41 to .43). In Grade 4, these inter-correlations were even higher (.47 to .65).

7.3.2 Reliability

Retest reliability was analysed for a subsample of 169 students. As shown in Table 6, for Subtest I, lower values of retest-reliability (r_{tt}=.58) were determined than for Subtest II (r_{tt}=.73). The students scored significantly higher at time 2 compared to time 1, both in Subtest I (T(168)=30.66, p<.01) and in Subtest II (T(168)=28.33, p<.01). The retest-reliability of Subtest II may thus be considered as acceptable (Bühner, 2011).

Table 6: Retest-reliability

Grade	Subtest I		Correlation coefficient	Subtest II		Correlation coefficient
	Time 1 ***M (SD)***	**Time 2** ***M (SD)***	***r***	**Time 1** ***M (SD)***	**Time 2** ***M (SD)***	***r***
3	6.56 (2.86)	8.08 (3.14)	.52**	6.40 (2.78)	8.12 (3.02)	.71**
4	9.25 (3.00)	10.28 (3.35)	.53**	9.54 (3.64)	10.86 (4.75)	.68**
total	7.47 (3.17)	8.82 (3.37)	.58**	7.46 (3.42)	9.04 (3.9)	.73**

Note: ** *p*<.01

8. Discussion

The present paper has discussed the construction of two GraLeV subtests and associated testing for usability. The paper has also reported upon two studies on piloting and the determination of the test quality criteria (reliability and validity). Both Subtest I and Subtest II were constructed to assess text level reading comprehension. In Subtest I, students read short nonsense-stories and answered two corresponding questions (one extracting explicit information and one making inferences). Subtest II is based on the maze procedure. Students are presented with texts in which every seventh word is replaced by a drop down in which students need to pick the word that best fits the context of the sentence in each case.

Study 1 reported on the piloting of the subtests (N=273 Grade 3 and 4 students) to gain information on item characteristics and test quality (internal consistency as reliability). Internal consistency of the final version of Subtests I was found to be satisfactory (α=.87), and a time limit of 3 minutes was set. The internal consistency of Subtest II was satisfactory for both texts (α=.78 to .80). Due to the speed test character of this subtest, the data from Study 2 could not be used to confirm internal consistency. As we wanted to enhance readability on a tablet screen, we decided against a suggested time limit of three minutes (Conoyer et al., 2017), and set the time limit at 100 seconds. However, by using two texts (one relatively easy, and one relatively demanding), we were able to differentiate assessment in terms of slow and fast readers.

Study 2 analysed test quality of both subtests. The results showed that both subtests fulfil the quality criteria of validity, showing high correlations with the ELFE II reading comprehension test and also with TA of text comprehension, but exhibit lower correlations with TA of mathematical or socio-emotional skills. TAs were also found to correlate highly with each other. Studies have shown that teachers tend to have a rather holistic picture of their students and do not differentiate between various sub-abilities (Paleczek, Seifert and Gasteiger-Klicpera, 2015). This also explains the higher correlations between GraLeV and ELFE II than those between GraLeV and TA of reading skills, as well as the significant inter-correlations between the different TAs. This result highlights the importance of regularly using standardised tests in order to aid teachers in identifying students with reading deficiencies.

Both Subtests I and Subtest II assess text comprehension to a certain extent. However, as can be seen in the correlation values for the ELFE II subtests, both subtests also capture sentence comprehension to a high degree. As the hierarchical reading levels are highly intertwined (Richter and Christmann, 2009), this result is not surprising. Future research will examine whether the implementation of both text level subtests really brings a greater gain in knowledge or whether one subtest at the text level is sufficient.

Retest reliability was found to be satisfactory for Subtest II (r_{tt}=.73), but not for Subtest I (r_{tt}=.58). Analysis also revealed that students scored higher at time 2 than at time 1. This is probably due to the unfamiliarity of the test at time 1. At time 2, students had accustomed themselves to the digital test environment, knew what they had to press, and were therefore faster, i.e. they were able to solve more items in the given time. Particularly in Subtest I, where each item was presented on a separate tablet screen, the time required for items may have acted as a break on students with relatively little digital experience, and thus have led to fewer items being solved. Although usability was evaluated beforehand, as some students might have had only limited prior digital experience (Medienpädagogischer Forschungsverband Südwest, 2018), differences between the students might not have been caused by differences in reading comprehension but by differences in digital competence. This needs to be considered in future studies of digital testing.

Study results suggest that both GraLeV subtests can be used to reliably and validly assess reading comprehension digitally. The manifold advantages of digital tools in learning and assessment (e.g., economic and objective administration, improved scoring and interpretation process; Neumann et al., 2019) justify further developments and evaluations of digital assessments. However, print assessments will remain important for certain classrooms or particular students. For example, internet problems might make it necessary to use the GraLeV in print form. It is thus important to investigate the potential equivalence of the GraLeV print and digital versions in future. Moreover, we are currently developing an app for tablet use that enables us to conduct the GraLeV without relying on the internet. Internet is only needed to upload data, thus enhancing practicability in classroom settings. Retest-reliability can then be analysed again and we are confident that this will lead to better results for Subtest I, where in the current internet-dependent version, the loading of each question on a new screen may have caused delays in students' answers. Standardization of the GraLeV will take place in autumn 2021. The test will then be freely available for teachers. Thus, the GraLeV will soon provide primary school teachers with a valuable digital tool for (a) simplifying differentiated assessment of student text comprehension and (b) easing documentation of learning progress.
Please add:

Acknowledgements

The authors acknowledge the financial support by the Univeristy of Graz and the support by the interns in the project RegioDiff.

References

Artelt, C. et al., 2007. *Förderung von Lesekompetenz. Expertise*, Bundesministerium für Bildung und Forschung, Berlin.

Artelt, C. and Gräsel, C., 2009. "Diagnostische Kompetenz von Lehrkräften", *Zeitschrift für Pädagogische Psychologie*, Vol. 23, No. 34, pp.157–160. 10.1024/1010-0652.23.34.157.

Bamberger, R. and Vanecek, E., 1984. *Lesen-Verstehen-Lernen-Schreiben. Die Schwierigkeitsstufen von Texten in deutscher Sprache*, Jugend und Volk, Wien.

Blumenthal, S. and Blumenthal, Y., 2020. "Tablet or Paper and Pen? Examining Mode Effects on German Elementary School Students' Computation Skills with Curriculum-Based Measurements", *International Journal of Educational Methodology*, Vol. 6, No. 4, pp.669–680. https://doi.org/10.12973/ijem.6.4.669.

Brandhofer, G., 2015. *Die Kompetenzen der Lehrenden an Schulen im Umgang mit digitalen Medien und die Wechselwirkungen zwischen Lehrtheorien und mediendidaktischem Handeln*, unpublished dissertation, Technical University Dresden.

Bühner, M., 2011. *Einführung in die Test- und Fragebogenkonstruktion*, Pearson Studium, München.

Chall, J.S., 1983. *Stages of Reading Development*, McGrawBHill Book Company, New York.

Cheung, A.C.K. and Slavin, R.E., 2012. "How features of educational technology applications affect student reading outcomes: A meta-analysis", *Educational Research Review*, Vol. 7, pp.198-215.

Conoyer, S.J., Lembke, E.S., Hosp, J.L., Espin, C.A., Hosp, M.K. and Poch, A.L., 2017. "Getting More From Your maze: Examining Differences in Distractors", *Reading & Writing Quarterly,* Vol. 33, No. 2, pp.141-154. 10.1080/10573569.2016.1142913.

Ehlers, J.P., Guetl, C., Höntzsch, S., Usener, C.A. and Gruttmann, S., 2013. "Prüfen mit Computer und Internet. Didaktik, Methodik und Organisation von E-Assessment" in Ebner, M. and Schön, S. (eds.) *L3T – Lehrbuch für Lernen und Lehren mit Technologien*. [e-book] Available at <http://l3t.eu/homepage/> [Accessed 15 August 2018].

Eid, M. and Schmidt, K., 2014. *Testtheorie und Testkonstruktion*, Hogrefe, Göttingen.

Espin, C., Wallace, T., Lembke, E., Campbell, H. and Long, J. D., 2010. "Creating a progress-monitoring system in reading for middle-school students: Tracking progress toward meeting high-stakes standards", *Learning Disability Research & Practice*, Vol. 25, No. 2, pp.60–75. 10.1111/j.1540–5826.2010.00304.

Medienpädagogischer Forschungsverbund Südwest, 2018. KIM-Studie 2018. Basisstudie zum Medienumgang 6- bis 13-Jähriger in Deutschland. Available at <https://www.mpfs.de/fileadmin/files/Studien/KIM/2018/KIM-Studie_2018_web.pdf> [Accessed 10 03 2021].

Förster, N. and Souvignier, E., 2014. "Learning progress assessment and goal setting: Effects on reading achievement, reading motivation and reading self-concept", *Learning and Instruction*, Vol. 32, pp.91–100. https://doi.org/10.1016/j.learninstruc.2014.02.002.

Fuchs, L.S. and Fuchs, D., 1992. "Identifying a measure for monitoring student reading progress", *School Psychology Review,* Vol. 21, No. 1, pp.45-58. 10.1016/j.learninstruc.2014.02.002.

Gebhardt, M., Diehl, K. and Mühling, A., 2016. "Online Lernverlaufsmessung für alle SchülerInnen in inklusiven Klassen", *Zeitschrift für Heilpädagogik*, Vol. 66, pp.444-453.

Gellert, A.S. and Elbro, C., 2012. "Cloze Tests May be Quick, But Are They Dirty? Development and Preliminary Validation of a Cloze Test of Reading Comprehension", *Journal of Psychoeducational Assessment,* Vol. 31, No. 1, pp.16-28. 10.1177/0734282912451971.

Graesser, A.C., Singer, M. and Trabasso, T., 1994. "Constructing inferences during narrative text comprehension", *Psychological Review,* Vol. 101, No. 3, pp.371–395. 10.1037/0033-295X.101.3.371.

Guthrie, J.T., Seifert, M., Burnham, N.A. and Caplan, R.I., 1974. "The maze Technique to Assess, Monitor Reading Comprehension", *The Reading Teacher,* Vol. 28, No. 2, pp.161-168.

Ketterlin-Geller, L.R., McCoy, J.D., Twyman, T. and Tindal, G., 2006. "Using a Concept maze to Assess Student Understanding of Secondary-Level Content", *Assessment for Effective Intervention, Vol. 31, No.* 2, pp.39-50. 10.1177/073724770603100204.

Klicpera, C., Schabmann, A., Gasteiger-Klicpera, B. and Schmidt, B., 2017. *Legasthenie – LRS. Modelle, Diagnose, Therapie und Förderung*, 5th edn, UTB, Stuttgart.

Lehmann, R.H., Peek, R. and Poerschke, J., 2006. *Hamburger Lesetest für 3. Und 4. Klassen: HAMLET 3-4*, 2nd edn, Hogrefe, Göttingen.

Lenhard, W., 2013. *Leseverständnis und Lesekompetenz: Grundlagen – Diagnostik – Förderung,* Kohlhammer, Stuttgart.

Lenhard, W., Lenhard, A. and Schneider, W., 2020. *Ein Leseverständnistest für Erst- bis Siebtklässler – Version II: ELFE II*, 4th edn, Hogrefe, Göttingen.

Lenhard, W., Schroeders, U. and Lenhard, A., 2017. Equivalence of Screen Versus Print Reading Comprehension Depends on Task Complexity and Proficiency. *Discourse Processes*, Vol. 54, No. 5-6, pp.427–445. https://doi.org/10.1080/0163853X.2017.1319653

Mayringer, H. and Wimmer, H., 2014. *Salzburger Lese-Screening für die Schulstufen 2-9: SLS 2-9*, Hogrefe, Göttingen. 10.1080/0163853X.2017.1319653.

Moll, C. and Landerl, K., 2010. *Salzburger Lese- und Rechtschreibtest II: SLRT II*, Hogrefe, Göttingen.

Muijselaar, M.M.L., Kendeou, P., de Jong, P.F. and van den Broek, P.W., 2017. "What Does the CBM-maze Test Measure?", *Scientific Studies of Reading,* Vol. 21, No. 2, pp.120-132. 10.1080/10888438.2016.1263994.

Mühling, A. and Gebhardt, M., 2021. LEVUMI. www.levumi.de

Neumann, M.M., Anthony, J.L., Erazo, N.A. and Neumann, D.L., 2019. "Assessment and Technology: Mapping Future Directions in the Early Childhood Classroom". *Frontiers in Education*, Vol. 4, No. 116, pp.1–13. https://doi.org/10.3389/feduc.2019.00116

Nielsen, J., 1993. *Usability Engineering*, Academic press, Boston. 10.1080/0163853X.2017.1319653.

Paleczek, L. and Seifert, S., 2019. "Pädagogische Diagnostik und deren Bedeutung für inklusiven Leseunterricht" in Paleczek, L. and Seifert, S. (eds.) *Inklusive(r) Leseunterricht: Leseentwicklung, Diagnostik und Konzepte,* Wiesbaden, Springer VS, pp.125-147.

Paleczek, L., Seifert, S., Franz, A., Wohlhart, D. and Riedl, S. (in prep.). *Grazer Leseverständnistest: GraLeV*.

Paleczek, L., Seifert, S. and Gasteiger-Klicpera, B., 2017. "Influences on teachers' judgment accuracy of reading abilities on second and third grade students: a multilevel analysis", *Psychology in the Schools*, Vol. 54, No. 3, pp.228-245. https://doi.org/10.1002/pits.21993

Paleczek, L., Seifert, S. and Schöfl, M. (2021). Comparing digital to print assessment of receptive vocabulary with GraWo-KiGa in Austrian kindergarten. British Journal of Educational Technology, 52, 2145-2161. DOI: 10.1111/ bjet .13163

Richter, T. and Christmann, U., 2009. "Lesekompetenz: Prozessebenen und interindividuelle Unterschiede" in Groeben, N. and Hurrelmann, B. (eds.) *Lesekompetenz: Bedingungen, Dimensionen, Funktionen,* 3rd edn, Beltz, Weinheim.

Richter, T., Naumann, J., Isberner, M., Neeb, Y. and Knoepke, J., 2017. *ProDi-L: Prozessbezogene Diagnostik von Lesefähigkeiten im Grundschulalter*, Hogrefe, Göttingen.

Schaumburg, H., 2015. "Chancen und Risiken digitaler Medien in der Schule" in Bertelsmann Stiftung (eds) *Individuell fördern mit digitalen Medien. Chancen, Risiken, Erfolgsfaktoren*, Bielefeld, Verlag Bertelsmann Stiftung, pp.20-94.

Schnotz, W., 1994. *Aufbau von Wissensstrukturen: Untersuchungen zur Kohärenzbildung beim Wissenserwerb mit Texten,* Psychologie Verlags Union, Weinheim.

Südkamp, A., Kaiser, J. and Möller, J., 2012. "Accuracy of teachers' judgments of students' academic achievement: A meta-analysis", *Journal of Educational Psychology*, Vol. 104, No. 3, pp.743–762. 10.1037/a0027627.

Walter, J., 2013. *Verlaufsdiagnostik sinnerfassenden Lesens: VSL*, Hogrefe, Göttingen.

Wayman, M.M., Wallace, T., Wiley, H.Y., Tichá, R. and Espin, C.A., 2007. "Literature Synthesis on Curriculum-Based Measurement in Reading", *Journal of Special Education,* Vol. 41, No. 2, pp.85-120. 10.1177/00224669070410020401.

Widauer, K. and Wallner-Paschon, C., 2017. "Entwicklung und Aufbau der Testinstrumente und Kontextfragebögen" in Wallner-Paschon, C. and Itzlinger-Bruneforth, U. (eds.) *PIRLS 2016. Technischer Bericht*, Salzburg, pp.9-22.

Wild, J. and Pissarek, M., 2019. *Regensburger Analysetool für Texte. Dokumentation*, available at:<https://www.uni-regensburg.de/sprache-literatur-kultur/germanistik-did/downloads/ratte/index.html> [Accessed 25 July 2019].

Wild, J. and Pissarek, M., n.d. *Ratte. Regensburger Analysetool für Texte*, available at:<http://www.uni-regensburg.de/sprache-literatur-kultur/germanistik-did/ratte/index.html> [Accessed 28 August 2019].