Spelling Error Detection: A Valid and Economical Task for Assessing Spelling Skills in Elementary School Children

Darius Endlich[1], Tobias Richter[1], Peter Marx[1], Wolfgang Lenhard[1], Kristina Moll[2], Björn Witzel[2] and Gerd Schulte-Körne[2]

[1] Department of Psychology IV, Educational Psychology, University of Würzburg

[2] Department of Child and Adolescent Psychiatry, Psychosomatics, and Psychotherapy, University of Munich

Accepted for publication in

***Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie* (2021)**

Korrespondierender Autor:

Dr. Darius Endlich, Universität Würzburg, Lehrstuhl für Psychologie IV, Röntgenring 10, 97070 Würzburg, darius.endlich@uni-wuerzburg.de

Zusammenfassung

Rechtschreibung zählt zu den Schlüsselkompetenzen für schulischen und beruflichen Erfolg. Um Kinder mit Rechtschreibproblemen adäquat zu unterstützen, ist eine frühe, möglichst niederschwellige Diagnostik essentiell. Aufgaben, in denen Rechtschreibfehler in präsentierten Texten zu identifizieren sind, könnten derartige ökonomische Verfahren darstellen. Obgleich Fehleridentifikationstests im angloamerikanischen Sprachraum weit verbreitet sind, haben sich die wenigen Studien im deutschsprachigen Raum bisher ausschließlich mit Kindern der Sekundarstufe beschäftigt. Die vorliegende Arbeit untersuchte in vier unabhängigen Studien $N = 1.513$ Grundschulkinder. Mittels linearer Regressionen wurden Rechtschreibkompetenzen (erhoben durch Fließ- und Lückendiktate) durch Leistungen in Fehleridentifikationstests vorhergesagt. Leistungen im Fehleridentifikationstest sagten Rechtschreibkompetenzen in allen Studien signifikant voraus (R^2 zwischen .509 und .679), was eine starke Assoziation der beiden Maße belegt. Prädiktive Werte zur Identifikation von Kindern mit schwachen Rechtschreibleistungen durch den Fehleridentifikationstest waren gut. Fehleridentifikation als Maß für Rechtschreibkompetenzen ist damit ein valides Instrument nicht nur für den angloamerikanischen Sprachraum, sondern auch für transparente Sprachen.

Schlüsselwörter: Rechtschreibung, Diktat, Fehleridentifikation, Lese-Rechtschreibstörung, Diagnose

Abstract

The ability to spell words correctly is a key competence for educational and professional achievement. Economical procedures are essential to identify children with spelling problems as early as possible. Given the strong evidence showing that reading and spelling are based on the same orthographic knowledge, error detection tasks (EDT) could be considered such an economical procedure. Although EDT are widely used in English-speaking countries, the few studies in German-speaking countries only investigated pupils in secondary school. The present study investigated $N = 1.513$ children in elementary school. We predicted spelling competencies (measured by dictation or gap-fill dictation) based on an EDT via linear regression. Error detection abilities significantly predicted spelling competencies (R^2 between .509 and .679), indicating a strong connection. Predictive values in identifying children with poor spelling abilities with an EDT were sufficient. Error detection for the assessment of spelling skills therefore is a valid instrument for transparent languages as well.

Keywords: spelling, dictation, error detection, developmental dyslexia, diagnosis

Spelling Error Detection: A Valid and Economical Task for Assessing Spelling Skills in Elementary School Children

The ability to spell words correctly is a basic skill not only for academic achievement but also for occupational careers and many everyday activities in modern information societies. Children with low spelling competencies have an increased risk of school failure (Jimerson, Egeland, Sroufe, & Carlson, 2000). Spelling competencies are very important for school and working life. They might even have a stronger impact than intelligence on decisions regarding the choice of the subsequent educational track (see W. Schneider, 2008a). National educational standards (e.g., in Germany) commonly include spelling as a core competency to be learned in elementary school (KMK, 2005). Difficulties in reading and writing occur frequently in the first two years of elementary school, and performances in spelling seem to be quite stable throughout elementary school (Klicpera & Gasteiger-Klicpera, 1998). Therefore, diagnosis and supporting measures for children with low spelling competencies should be considered as early as possible (Mannhaupt, 1994).

Currently, there is a lack of economical screening procedures for spelling skills in German elementary school children. The common way to assess spelling abilities are gap-fill dictation tasks. These tasks require time and other resources to administer and code. Moreover, these tasks cannot be implemented as computerized tests to elementary school pupils because of the large individual differences in typing skills, which are mainly lacking (Jiménez, Marco, Suárez, & González, 2017).

Error detection may seem an unusual task to measure spelling skills because the task involves reading with no writing. However, we argue that theories of reading and spelling imply that both abilities are based on the same orthographic knowledge. Moreover, we review evidence suggesting a close convergence of spelling assessments based on error detection and on dictation. However, this evidence comes from English-speaking countries

or only assessments designed for secondary school students in other countries. The present study is the first empirical study that has examined the validity of error detection tasks for assessing spelling abilities in German-speaking elementary school children.

Dual-Route Model of Reading and Spelling

The dual-route model of reading (Coltheart, 1978) postulates a lexical and a nonlexical route to recognize written words. The lexical route is characterized by a direct access to a mental lexicon through orthographical representations of word forms, whereas using the nonlexical route involves recoding the printed word letter by letter and translating the graphemes into phonological code based on grapheme-phoneme correspondence rules (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001).

A similar dual-route model of spelling is the standard model for describing and distinguishing different types of spelling disorders in children (Houghton & Zorzi, 2003; Romani, Olson, & Di Betta, 2005, p. 432). Similar to the dual-route model of reading, such models assume two possible ways of spelling words correctly. The first route leads from the phonological lexicon (storing the sound of the word when hearing it) and the semantic system (storing the meaning of the word) to the orthographic lexicon (storing word spelling in the form of series of abstract letter entities). The second, nonlexical-route is analogous to the indirect way in the dual-route model of reading, with the only difference that phoneme-grapheme correspondences are needed to recode phonemes into graphemes. Many of the processes and representations that are involved in spelling are also involved in reading, although the process runs in reverse from sounds to letters rather than from letters to sounds (Romani et al., 2005, p. 434; for a unified dual-route model of reading and spelling, see Rapcsak, Henry, Teague, Carnahan, & Beeson, 2007).

Some strong evidence exists showing that reading and spelling are based on the same orthographic knowledge. Based on their studies on spelling low-frequency words, Burt and

Tate (2002) concluded that a single orthographic lexicon serves visual word recognition and spelling. Holmes and Babauta (2005) reported that students could rarely improve their spelling after writing known words with minimal visual feedback beforehand. The authors concluded that individuals acquire a single orthographic representation from repeated exposures to a word from reading and spelling. However, recognition procedures such as error detection tasks may not be sufficient to predict spelling competencies. Although reading and spelling share a common representation, some representations may be incomplete (Holmes & Carruthers, 1998). Therefore, reading and spelling cannot be seen simply as inverse processes. Partial cues could allow readers to identify the target word even though they are not able to spell the same word correctly.

Measurement of Spelling Competencies and Validity of Error Detection Tasks

Additional evidence for a close connection of reading and spelling comes from studies that examined the relationships of different spelling tasks. There are various ways to measure spelling competencies. In Germany, dictations are still one of the most common ways to assess spelling in elementary school, despite serious criticism. One criticism is that dictations cannot be fair for slow- and fast-writing children in one class (Staatsinstitut für Schulqualität und Bildungsforschung [State Institute for School Quality and Educational Research], 2005). Moreover, evaluating dictations is often not objective. In a study of 415 teachers evaluating the same dictations, Birkel (2009) found a high spread of identified misspelled words and therefore a low interrater reliability of grades (number of words marked as misspelled varied up to 41 across teachers).

Besides the active component of spelling measured by dictations or gap-fill dictations, the competency to identify misspelled words was included in the German educational standards (KMK, 2005). Many studies suggest that the active and passive component of spelling abilities are highly associated, despite their slightly different cognitive

requirements. Error detection and error correction tests are commonly used in English-speaking countries. However, only three such tests exist for the German language (HSP 5-10 EK, May, 2012; R-FIT 5-6+, M. Schneider, Martinez Méndez, & Hasselhorn, 2014; R-FIT 9-10, Lenhart, Marx, Segerer, & W. Schneider, 2019). All of these tests are designed for Grade 5 upwards. In the R-FIT 5-6+ and R-FIT 9-10 (Fehleridentifikationstest – Rechtschreibung für fünfte und sechste Klassen / für neunte und zehnte Klassen), the errors in misspelled words must be marked by a vertical line. Correlations between .81 and .83 of R-FIT 5-6+ with a standardized dictation task indicate a very good convergent validity for an error detection task in Grades 5-6. Similar results have been demonstrated for Grades 9 and 10. Lenhart et al. (2019) reported a correlation between R-FIT 9-10 and a gap-fill dictation task of .88. The RIOC index (Relative Improvement Over Chance; see Loeber & Dishion, 1983) for detecting children with below-average spelling skills (defined as the 16th percentile in the Grade norms for the dictation task) was 61% for R-FIT 5-6+ and 76% R-FIT 9-10, which also indicates a remarkable convergence of assessments based on error detection and dictation. In the HSP 5-10 EK (Hamburger Schreib-Probe – Erweiterte Kompetenzen; May, 2012), children's task is to correct misspelled words directly. To our knowledge, no information about convergent validity with dictation tasks is available for this test.

In English-speaking countries, many studies have examined proofreading as an economical procedure for assessing spelling skills. Nearly a century ago, Guiler (1929) compared different types of spelling test methods, for example, a multiple-choice test with one correctly spelled version out of four different incorrect spellings of a word. A factor analysis of Allen and Ager (1965) found no difference between tasks measuring the active and passive component of spelling, that is, scores obtained in both tasks were highly correlated and loaded on one single factor. Freyberg (1970) compared three different

spelling tasks (free writing, dictation and error detection multiple-choice-tests) for students from secondary school in New Zealand. Freyberg concluded that dictations (recall) and error detection (recognition) measured nearly the same competency given that the highest correlation was found between these two tasks. Finally, strong correlations between error detection tasks and dictations have been found also in elementary school for Grades 1 to 6 (Allred, 1984), Grades 2, 3, 5 and 7 (Frisbie & Cantor, 1995) and for Grades 2 to 5 (Westwood, 1999). Frisbie and Cantor (1995) recommended a multiple-choice task with four different words and the possibility that one or none of the words were misspelled.

To conclude, high correlations between error detection tasks and dictations are reported for elementary school in English-speaking countries (r = .59 – .90). In German-speaking countries, consistently high correlations have been found for Grades 5 and 6 (r = .81 – .83; M. Schneider et al., 2014) and Grades 9 and 10 (r = .88; Lenhart et al., 2019). Yet, whether the findings obtained in secondary school students are transferable to elementary school in German-speaking countries remains an open question. Bearing in mind the importance of early diagnosis of spelling abilities and the identification of children with poor spelling competencies in particular, closing this gap would be an important advancement in the empirical literature on the topic of spelling and error detection in German elementary school children.

Research Rationale and Hypotheses

The aim of the present studies was to assess the connection between spelling and error detection tasks in German elementary school. To this end, we used a new tablet-based error detection test that is part of a comprehensive test battery for screening learning disorders (Endlich, Lenhard, Marx, & Richter, 2021; T. Richter, Lenhard, Marx, & Endlich, 2018). Given that the dual-route model approaches for reading and spelling suggest that reading and spelling are based on the same orthographic knowledge, we assumed that

spelling skills – measured by dictation tasks – could be reliably predicted by receptive error detection tasks.

Studies from English-speaking countries have consistently reported high correlations between spelling skills and performance in proofreading tasks from elementary to secondary school (Allen & Ager, 1965; Allred, 1984; Croft, 1982; Freyberg, 1970; Frisbie & Cantor, 1995; Westwood, 1999). The two available studies from German-speaking countries, both of which focus on secondary school children, yielded similar results (Lenhart et al., 2019; M. Schneider et al., 2014). Considering the high stability of spelling skills from elementary to secondary school in German-speaking school children (see W. Schneider, 2008b), we assumed a strong linear relationship between performance in error detection and dictation tasks for German elementary school children (Hypothesis 1). Considering the often reported advantages for girls in spelling skills in elementary school (S. Richter, 1994; Schneider, 1994; Schneider & Näslung, 1999), we added gender as predictor to all models. Hypothesis 2 focuses on children with poor spelling abilities. In analogy to similar findings in secondary school children (Lenhart et al., 2019; M. Schneider et al., 2014), we expected that elementary school children with poor spelling abilities (measured by various kinds of dictation tasks) could be reliably identified with an error detection task (RIOC index > .66). Considering the high correlations that are usually reported between spelling skills and performance in proofreading tasks and between two active spelling tasks (e.g. criterion validity in DERET3-4+ $r \geq .64$), we also tested whether a receptive error detection task would still predict spelling skills, as measured with a dictation task, even after adding another active spelling task (gap-fill dictation) to a multiple regression model (Hypothesis 3).

These hypotheses were tested with children in Grades 3 and 4. At this grade level, all graphemes and the alphabetic principle should have already been acquired, and reading and

writing instruction turns toward teaching orthographic rules and irregular deviations from the orthographic principles. We conducted four independent studies in different schools in Grade 3 and 4 (N = 1.513; for an overview, see Table 1) between July 2018 and August 2020. Studies 1, 2 and 3 were carried out in the children's normal classroom. Children only participated if their parents agreed by signing an informed consent document beforehand. Thus, 69.70% of the overall sample participated in our Studies 1, 2 and 3. The period of the testing was limited to one 45-min classroom lesson. To gain comparable group sizes, the maximum of participants in one classroom was set to 20 (M = 15.06, SD = 3.86, *Minimum* = 8, *Maximum* = 20). If there were more than 20 children with parent agreement in one class, those children were assigned to the testing session in another class. Thus, all children with parent agreement participated. Testing was conducted by trained student assistants. After a brief instruction, children started the error detection task (EDT) on a tablet. The second part of the testing was either a dictation or a gap-fill dictation task, both provided as paper-based tests.

In Study 4, data was collected through online testing. Beyond spelling tests, a reading fluency test was provided, which allowed investigating the discriminant validity of the EDT. Considering that fluent reading and spelling share a common knowledge base, we expected substantial correlations between error detection and reading fluency. However, considering the processual differences between reading and spelling, we expected the correlations of error detection and reading fluency to be lower than the correlations of error detection and active spelling (Hypothesis 4).

[Table 1 near here]

Study 1

Method

Participants. The sample of Study 1 were 98 students in Grade 3 and 4 recruited from one elementary school in Bavaria, Germany (Grade 3: $n = 45$, 52% female; Grade 4: $n = 53$, 58% female). The data collection was fully anonymized, including the age of the participant. However, in German primary schools, the age range of children in Grade 3 is 8-9 years and in Grade 4, 9-10 years, with typically little variation. Data collection took place in a period of two weeks at the end of the school year (July 2018).

Instruments. The first measure of spelling skills was a commonly used spelling test based on a dictation of a continuous text, Form A of the German spelling test for Grades 3 and 4 (Deutscher Rechtschreibtest für das dritte und vierte Schuljahr; DERET 3-4+; Stock & W. Schneider, 2008a). The test consists of a dictation with single sentences read repeatedly and one after the other (Grade 3: 80 words; Grade 4: 92 words). The number of words read aloud contained two to five words each. For the sample reported in the manual of the DERET, the internal consistency (Cronbach's $\alpha \geq .92$), the split-half reliability ($r \geq .90$) and test-retest reliability ($r \geq .81$) of this test were very good and the criterion validity was at least satisfying ($r \geq .64$ between DERET and DRT 3 (Müller, 1996) and DRT 4 (Grund, Haug, & Naumann, 1994). The test score is the number of misspelled words, ranging from 0 to 80 (Grade 3) and from 0 to 92 (Grade 4), with zero representing the best performance possible. The testing took 15 to 20 minutes.

The second measure of spelling skills was a newly constructed computerized test that was based on an error detection task (EDT) and presented on a computer tablet (10.1 inch). The students' task was to identify 86 misspelled words in continuous texts (284 words in sum). Instructions were presented via headphones and through visually displayed examples on the tablet screen. The task was split into three stories to maintain interest. At the beginning of each section, the whole text was read to the children via headphones. Then, every single sentence of the story was presented visually and auditorily. The task was to

identify the misspelled words (1 to 6 misspelled words per sentence; $M = 2.63$, $SD = 1.26$). Children were able to correct their responses as long as the sentence was displayed, but after navigating to the next sentence, there was no possibility to go back. The first text included 32 misspelled words (total: 91 words), the second text 26 (total: 96 words) and the third text 28 (total: 97 words). The score for the EDT was calculated as the difference of the number of misspelled words and the sum of the identified misspelled words plus false alarms (i.e., correctly spelled words marked as false). Thus, the EDT scores ranged from 0 (= best performance) to 284. Internal consistency reliability of the EDT was very good ($\omega_t = .95$; see McNeish, 2018). The test took approximately 15 to 20 min. In contrast to the dictation task, performance indicators (accuracy on single word level and response latency on sentence level) were automatically recorded (mean response latency in Grade 3: $M = 10.51$ s, $SD = 4.24$ s; Grade 4: $M = 10.05$ s, $SD = 1.84$ s).

We constructed the EDT following the lead of error detection tasks developed for secondary school (e.g., M. Schneider et al., 2014). Spelling errors were inserted by omitting or adding letters, replacing letters with wrong letters and case errors. All the misspelled words were taken from the core vocabulary of primary school children and validated with the online data base childLex (Schroeder, Würzner, Heister, Geyken, & Kliegl, 2014). Furthermore, we created two different forms of misspelled words: (1) phonologically correct misspellings (e.g., *unt* for *und* [and]; equivalent to alphabetic or orthographic stage, Thomé & Thomé, 2017) and (2) phonologically incorrect misspellings (e.g., *Ausflaug* for *Ausflug* [trip]). Errors of category (1) were noticeably more difficult than errors of category (2), both for children in Grade 3 ($M = 62\%$ correct vs. $M = 83\%$ correct) and in Grade 4 ($M = 80\%$ correct vs. $M = 89\%$ correct). Given that the EDT was designed as a screening instrument, the mistakes were explicit and easily recognizable for the average adult reader. The misspellings in the EDT were randomly distributed across the texts.

Statistical Analysis and Missing Data. The data from two students (one student in Grade 3 and one student in Grade 4) were not recorded because they aborted the app, and one student had too many missing words in the continuous dictation. Since missing data affected very few participants (0.8% resp. 1.2%), these three students were excluded from the analysis (listwise deletion, Enders, 2010, p. 39). ICCs for the dependent variable were lower than .05 both for Grade 3 and Grade 4, indicating that clustering effects were low in the present sample. Therefore, we decided to proceed with ordinary least squares (one-level) regression models. Given the often-reported advantages for girls in spelling skills in elementary school (S. Richter, 1994; Schneider, 1994; Schneider & Näslund, 1999), we added gender to all models.

To test Hypothesis 2, participants were divided into two groups. Children with poor spelling abilities (1 *SD* below the mean of the DERET raw score by grade level) and children with average or above-average spelling abilities. Children scoring in the 25th percentile in the error detection task (predictor variable) were classified as children at risk. Given that the EDT was designed as a screening instrument, we increased the percentage of children classified as children at risk, accepting a higher percentage of false alarms in order to reduce false negatives. Then, sensitivity (percentage of children with poor spelling abilities classified as children at risk), specificity (percentage of children with uncritical spelling abilities classified as children without risk) and the RIOC index were computed for Grades 3 and 4.

Results

Fourth graders performed noticeably better than third graders in the EDT (Table 2). Children at the end of Grade 4 overlooked only 16 misspelled words ($M = 16.25$, $SD = 10.07$), whereas children at the end of Grade 3 overlooked about twice as many ($M = 28.82$, $SD = 14.71$), $t(93) = -4.91$, $p < .001$, $d = -1.01$.

[Table 2 near here]

The correlation coefficient between the EDT score and its main component, the sum of identified misspelled words, was extremely high (r = .99). False alarms occurred seldom and were basically uncorrelated with the EDT score and other measures.

Hypothesis 1. Multiple linear regression models were estimated to predict the DERET (T score) with the EDT score and gender as predictors (Table 3). The model explained a significant and considerable proportion of variance in the DERET scores in Grade 3, R^2 = .526, $F(2,41) = 22.75$, $p < .001$, and in Grade 4, R^2 = .581, $F(2,48) = 33.29$, $p < .001$ (Figure 1). Only the effect of the EDT was significant in both models. Thus, Hypothesis 1 was supported.

[Figure 1 near here]

[Table 3 near here]

Hypothesis 2. When the EDT scores were used to predict children with poor spelling abilities (determined with the dictation-based DERET), good sensitivity (SN) and specificity (SP) were obtained for both Grade 3 (SN = 100.0%; SP = 74.4%) and Grade 4 (SN = 80.0%; SP = 78.3%). The RIOC indices were high (Grade 3: 100.0%; Grade 4: 72.4%). In line with Hypothesis 2, error detection seems to be an appropriate task in Grades 3 and 4 for identifying children with poor spelling abilities (SN = 90.0%; SP = 76.5%; RIOC = 85.6%; Table 4).

[Table 4 near here]

Study 2

Method

Participants. The sample of Study 2 were 107 students in Grades 3 and 4, again recruited from one elementary school in Bavaria, Germany (Grade 3: n = 52, 52% female; Grade 4: n = 55, 53% female). In contrast to Study 1, time of data collection was about three

months after the beginning of the school year (December 2018). Therefore, spelling performance between Study 1 and 2 is not directly comparable.

Design, Instruments and Missing Data. Test design was very similar to Study 1. The error detection task was the same as in Study 1, with one exception. The single sentences were presented only visually. Internal consistency reliability of the EDT again was very good (McDonald's $\omega = .95$). In contrast to Study 1, a gap-fill dictation of the DERET was provided. In Grade 3, the gap-fill dictation for the second grade was administered (DERET 1-2; Stock & W. Schneider, 2008b) and for Grade 4, the gap-fill dictation for the third grade (DERET 3-4). Participants completed the task for the previous grade level because the tests are designed to be administered at the end of the school year or at the beginning of the next school year. Sixteen sentences (Grade 3) or 14 sentences (Grade 4), with one to three gaps in each sentence, were read aloud by the test administrator. As in Study 1, the test score was the number of misspelled words, ranging from 0 to 24 (Grade 3) and from 0 to 28 (Grade 4), with zero representing the best performance possible. The test took approximately 15 min to complete.

No missing data were found. The data analysis strategy was nearly the same as in Study 1. The gap-fill dictation tasks varied for grade level and no norm scores for the gap-fill dictation were provided in the DERET manual. Thus, DERET raw scores were transformed into *z* values by grade level for building the extreme group split.

ICC for the dependent variable was lower than .05 for Grade 3 and .06 for Grade 4, again indicating that clustering effects due to classes were not an issue.

Results

Three months after the beginning of the school year, fourth graders overlooked about 34 misspelled words in the error detection task. Therefore, their spelling performance was worse than third graders at the end of the school year ($M = 28.82$; see Study 1), which might

indicate an additional handicap in the EDT version without the additional reading-aloud feature. False alarms seem to play a role as indicated by the significant correlation coefficient of .36 between false alarms and the sum of identified misspelled words. Comparable to Study 1, fourth graders performed noticeably better than third graders in the EDT. Descriptive data and correlations are provided in Table 2.

Hypothesis 1. Multiple linear regression models were estimated to predict the DERET score (*z* score), with the EDT score and gender as predictors (Table 3). The model explained a significant and considerable proportion of variance in the DERET scores in Grade 3, $R^2 = .611$, $F(2,49) = 38.55$, $p < .001$, and in Grade 4, $R^2 = .679$, $F(2,52) = 54.95$, $p < .001$ (Figure 1). Only the effect of the EDT score was significant in both models.

Hypothesis 2. Predictive values for gap-fill dictation were slightly weaker for Grade 3 (SN = 80.0%; SP = 88.1%; RIOC = 73.3%) and for Grade 4 (SN = 66.7%; SP = 84.8%; RIOC = 56.3%) compared to the predictive values obtained for the criterial dictation task used in Study 1. Nonetheless, to predict poor spelling abilities in gap-fill dictation tasks, error detection seems to be an appropriate task in Grades 3 and 4 (SN = 73.7%; SP = 85.2%; RIOC = 64.8%).

Study 3

Method

Participants. The sample of Study 3 were 54 students in Grade 4 (German elementary school; 43% female). Data collection took place about four months after beginning of the school year (January 2019).

Design, Instruments and Missing Data. Test design was very similar to Study 1. In addition to the EDT (same version as in Study 1) and DERET 3-4 (dictation for the third grade), a gap-fill dictation was provided using the same words as in one text of the EDT. A total of 26 words had to be filled in, in 11 sentences of one of the texts (Ein Besuch im Zoo

[A visit to the zoo]). Again, the test score was the number of misspelled words, ranging from 0 to 26.

Internal consistency reliability of the EDT again was very good (McDonald's ω = .95). No missing data was found. Two hierarchical linear regression models were estimated to predict the DERET score (*z* score), with the EDT score and the gap-fill dictation as predictors (Table 5; Figure 2).

[Figure 2 near here]

[Table 5 near here]

The ICC for the gap-fill dictation was lower than .05 and the ICC for DERET score was .11. However, the latter ICC was not significantly different from 0 ($z = 0.79$, $p = .429$), suggesting that clustering effects due to classes were not an issue in Study 3.

Results

Four months after beginning of the school year, fourth graders overlooked about 24 misspelled words on average in the error detection task. Therefore, their mean performance was between the third graders and the fourth graders at the end of the school year, which was comparable with the error detection skills of children at the end of Grade 3 (see Study 1). Notably, the correlation coefficient between the two different types of active spelling performance—continuous dictation and gap-fill dictation—was high ($r = .68$, $p < .001$) but not higher than the correlation between an active spelling and an error detection task ($r = .78$ for continuous dictation; $r = .67$ for gap-fill dictation; comparison of correlation coefficients of dependent samples; Steiger, 1980). Descriptive data and correlations are provided in Table 6.

[Table 6 near here]

Hypothesis 1. Two hierarchical linear regressions were estimated to predict the DERET score (*z* score) with the gap-fill dictation scores as predictor (Model 1), the EDT

score as predictor (Model 2) and both predictor variables (Model 3). All three models explained significant and considerable proportions of variance in the DERET scores. Model 2 had a slightly better fit (R^2 = .601) than Model 1 (R^2 = .467). Moreover, when the predictor variable EDT was added to Model 1, the model fit improved significantly (ΔR^2 = .184).

Hypothesis 2. When the EDT scores were used to predict children with poor spelling abilities (determined with the dictation-based DERET), predictive values were good. Six out of seven children with a DERET *T* score lower than 40 were identified by the EDT (SN = 85.7%), whereas 42 of 47 children with a DERET *T* score of 40 or higher were classified as "no risk" by the EDT (SP = 89.4%). The RIOC index also indicated highly reliable classification (82.1%).

Hypothesis 3. When adding the gap-fill dictation scores to Model 2, the model fit improved but to a relatively small extent (ΔR^2 = .050). Importantly, the EDT remained a significant predictor after gap-fill dictation was included as predictor in the model (Model 3) and its regression coefficient even exceeded the one of gap-fill dictation. Thus, Hypothesis 3 was supported.

Study 4

Method

Participants. The sample of Study 4 were 1.254 students in Grade 3 and 4 recruited from elementary schools in Bavaria, Germany (Grade 3: *n* = 646, 48% female; Grade 4: *n* = 608, 51% female). Parents stated that for 951 children (75.84%) German was the only language spoken at home. For 265 children (21.13%) German and a second language was spoken at home, and for 38 children (3.03%) only a language other than German was spoken at home. For 1.099 children (87.64%), German was the first language. Out of the remaining 155 children with a different first language, 120 children (9.57%) had learned German before the age of 4 years. Only 35 children (2.79%) with a different first language had not

learned any German before the age of 4 years. As in Study 1, we collected the data at the end of the school year (July 2020).

Design and Instruments. The error detection task was the same as in Study 1, with one difference. Only two out of three texts were presented, including a total of 54 misspelled words (total: 193 words). In contrast to Studies 1, 2 and 3, data was collected online. An invitation to participate in the study was sent to 22.500 families in the greater Munich area, including 7.000 families with children probably in Grade 3 and 4 (= 30% of the age cohort; children born between October 1, 2009 and September 30, 2011). The corresponding registration offices selected the families to be invited at random. These families received information about the study by post and were invited to download the required program from the internet. Parents were instructed to support their children only if they did not understand the instructions. It was pointed out that parents should not help their children in answering the questions. Children worked by themselves at home on several tablet-based tasks on two following days. On day two, children worked through three tests in the following sequence: a reading fluency test (computerized version of the Würzburger Leise Leseprobe – Revision; WLLP–R; Schneider, Blanke, Faust, & Küspert, 2011), a spelling test (computerized version of the Weingartener Grundwortschatz Rechtschreib-Test für dritte und vierte Klassen; WRT 3+ resp. WRT 4+; Birkel, 2007) and finally the EDT. Internal consistency was very good, for WRT 3+ (ω_t = .94), for WRT 4+ (ω_t = .96) and for WLLP-R (Grade 3: ω_t = .95; Grade 4: ω_t = .94).

Missing Data. Three participants had to be excluded from data analysis due to implausible high scores in combination with an extremely high number of false alarms in the EDT (63, 91, and 131). Given that false alarms occurred only seldom in the total sample (*M* = 0.54, *SD* = 0.97, *Minimum* = 0, *Maximum* = 17), these three children probably did not comply with the instruction.

Results

Fourth graders again performed noticeably better than third graders in the EDT (Table 2). Whereas the performance of the fourth graders in Study 4 was comparable to that of the fourth graders in Study 1 (83.0% vs. 81.1% of the misspelled words were identified), third graders performed significantly better in Study 4 (75.5%) than in Study 1 (68.4%), $t(651) = -2.97$, $p = .003$, $d = 0.47$. Again, the number of false alarms was basically uncorrelated with the EDT score and all other measures ($r < .20$). Mean response latencies on sentence level were slightly higher than in Study 1 for Grade 3 ($M = 12.11$ s; $SD = 5.36$ s), $t(651) = 1.97$, $p = .049$, and comparably high for Grade 4 ($M = 10.30$ s, $SD = 3.41$ s), $t(622) = 0.43$, $p = .67$.

Hypothesis 1. We used multiple linear regression models to predict the WRT (*T*-value) with the EDT score and gender as predictors (Table 7). The model explained a significant and considerable proportion of variance in the WRT scores in Grade 3, $R^2 = .588$, $F(2,643) = 459.5$, $p < .001$, and in Grade 4, $R^2 = .509$, $F(2,602) = 311.8$, $p < .001$). Only the effect of the EDT score was significant in both models.

[Table 7 near here]

Hypothesis 2. When the EDT scores were used to predict children with poor spelling abilities (determined with the dictation-based WRT), predictive values were comparable to those found in Study 2, both for Grade 3 (SN = 83.3%; SP = 82.4%) and for Grade 4 (SN = 83.1%; SP = 79.0%). The RIOC indices were high (Grade 3: 76.8%; Grade 4: 76.2%). In line with Hypothesis 2, the results suggest that error detection is an appropriate task in Grades 3 and 4 for identifying children with poor spelling abilities (SN = 83.2%; SP = 80.7%; RIOC = 76.6%).

Hypothesis 4: Correlations with Reading Fluency. In order to explore the discriminant validity, Pearson's correlations between dictation- and EDT-scores on the one

hand and reading fluency on the other hand were computed (Table 8). In line with Hypothesis 4, correlations between EDT and dictation were significantly higher in Grade 3 (r = .77) and in Grade 4 (r = .71) than between reading fluency and either of the orthographic measures (all correlations: $r \leq .57$; for all comparisons of correlations: $z > 4.090$, $p < .001$, Steiger's, 1980, test). Moreover, the correlation of reading fluency with productive spelling (WRT) was higher than the correlation with receptive spelling (EDT), both in Grade 3, $z = 3.53$, $p < .001$, and in Grade 4, $z = 2.37$, $p = .009$.

[Table 8 near here]

Discussion

The aim of the present studies was to investigate the validity of error detection tasks as a measure of spelling ability in elementary school. To this end, we closely examined the relationship between error-detection performance and the performance in active spelling tasks, that is, dictation and gap-fill dictation. In line with Hypothesis 1, error detection was closely related to active spelling tasks and explained more than 54% of the variance in all samples. Gender was not a significant predictor in any of the models. Performance differences between girls and boys were small descriptively and insignificant, in both active and in passive spelling tasks. Moreover, Study 3 indicated that error detection and dictation tasks show a comparably high or even higher correlation than gap-fill dictation and dictation tasks. The correlation between two active spelling tasks was not higher than the correlation between an active spelling and an error detection task. Likewise, in line with Hypothesis 3, error detection remained a significant and strong predictor of spelling performance in a dictation task, even after gap-fill dictation was added as predictor to a multiple regression model. This finding is particularly strong evidence for the validity of error detection tasks as a measure of spelling ability.

In line with Hypothesis 2, children with poor spelling abilities (criterion variable: performance in continuous or gap-fill dictation) could be reliably predicted by an error detection task. In sum, the results of all four studies testify that error-detection tasks are a valid and useful measure to assess spelling abilities in German third and fourth graders.

In order to prove the convergent and discriminant validity of the EDT, reading fluency was additionally measured in Study 4. Whereas productive and receptive spelling skills were highly correlated (≥ .70), correlations between reading fluency and the spelling skills were only moderate, thus supporting the convergent and discriminant validity of the EDT (Hypothesis 4). Furthermore, reading fluency was significantly higher associated with the productive spelling skills than with the receptive spelling skills. Although reading and spelling are based on the same orthographic knowledge (Burt & Tate, 2002; Holmes & Babauta, 2005), they cannot be seen simply as inverse processes. While sharing a common representation with reading, error detection tasks refer to spelling rather than to reading. To perform well in the EDT, children have to match the presented words with the corresponding entries in their orthographic lexicon. If the entry is missing, children have to consult knowledge about orthographic rules. Either way, the knowledge needed to spell words correctly – or to identify misspelled words – overlaps with but also exceeds the knowledge needed for fluent reading.

Establishing error detection tasks as standard instruments to measure spelling skills in elementary school would be beneficial in several ways. First, error detection tasks can be implemented as computer-based assessments, leading to highly feasible and economical testing procedures for spelling skills. Whereas evaluations of dictations take a lot of time and are prone to sources of error in applying and scoring the tests (Birkel, 2009), the evaluation of spelling skills measured by a computer-based error detection task is completed objective and readily available at the moment the student finishes the task. The EDT could therefore

be used not only for formative assessments but also to monitor the student's progress in times of homeschooling. The results on the development of spelling performance measured with the EDT across the school year (see Table 2) suggest that the EDT seems to be an appropriate instrument for repeated measures. Actually, the EDT in its current research version is to be implemented in a digital (web-based) program for identifying and promoting children with learning disabilities (LONDI – Lernstörungen; Online-Plattform für Diagnostik und Intervention, https://www.londi.de/). Moreover, measuring spelling skills with an error detection task might reduce test anxiety especially in children with poor spelling skills: They can work at their own pace and focus on every new sentence being presented. In dictations children with poor spelling abilities receive immediate feedback because they notice that they miss this word or another. In the EDT, on-task behavior is likely to be less compromised by expected test failure. During the test, children do not get immediate feedback on their performance and can thus concentrate on working on the task.

Limitations and Directions for Future Research

In the error detection task used in these studies, students are allowed to mark as many words per sentence as incorrect as they wish. Consequently, response tendencies and individual differences in careful reading could have contributed to the results. Some students may mark just one word to navigate very quickly to the next sentence, whereas other students read the whole sentence looking at every single word carefully. One possible solution could be to change the text materials and the instruction and ask students to mark exactly one word in every single sentence. Obviously, this task would differ from the current version of the error-detection task.

A second limitation is that in the present studies, order of presentation of the different tasks was not counterbalanced for practical reasons. Therefore, the presence of sequence and order effect, such as motivational effects, cannot be ruled out. For example, children in

Study 3 could have become tired when working on the gap-fill dictation after having completed the error detection and dictation tasks. Future studies should take care to control for sequence and order effects.

Third, despite the fact that the trend for digital testing is increasing and that investigations on the comparability of scores between paper- and computer-based tests have been present for more than three decades (Hassler Hallstedt & Ghaderi, 2018), a lack of knowledge still persists on the comparability of paper- and tablet-based tests for assessing spelling skills. For spelling tests, Berninger, Abbott, Augsburger, and Garcia (2009) reported differences between pen and keyboard writing that change from Grade 2 to Grade 6, possibly reflecting different developmental trajectories in the two modes of writing. For error recognition tests, in contrast, it seems plausible to assume that possible effects of presentation model (computer-/tablet-based presentation vs. paper-pencil-test) are much lower. Tablet-based error detection tasks do not require the test takers to be familiar with the keyboard. However, future studies should investigate the comparability of paper- and tablet-based error detection tasks.

Study 2 differs from the other studies regarding the assessment of the EDT. In analogy to error detection task for secondary school, where sentences had to be read by students themselves, we decided to present the sentences in Study 2 only visually, requiring students to decode the sentences on their own. This variation aimed at identifying the ideal setting for the application of EDTs in elementary schools. Although the change of the assessment strategy in one of the studies may be regarded as a limitation, it is also useful for exploring the robustness of the findings: Results were highly consistent across all four studies, regardless of whether the sentences were read out loud to the students or not.

Study 4 differs from the other studies regarding the study design as data was collected online. Variance in response latencies on sentence level was slightly lower when

children worked on the EDT in their classroom setting than working alone at home. This observation suggests that some of the children were more often distracted when working at home, leading to higher response latencies. Monitoring of student behavior is more limited in an online study compared to a classroom study when students' behavior can be observed directly. For example, we do not know whether and to what extent children were supported by parents or older siblings while working on the tablet-based task. Moreover, the sample of Study 4 might be less representative as only those children participated whose parents read and supported the invitation to the study. Furthermore, only children from families with a mobile device at home were able to participate. That said, Study 4 underscores the robustness of the findings obtained in Studies 1-3: The strong relationships between the error detection task and active spelling tasks were highly consistent across different study settings, in the classroom and at home.

In Germany, there are various curricula for teaching reading and spelling, for example, the spelling book approach (e.g. Metze, 2009; Bruhn et al., 2014) or the reading-through-writing approach (Lesen durch Schreiben, Reichen, 2006). These curricula greatly differ in their didactic approach to spelling instruction. For example, the reading-through-writing approach allows students to spell words in the orthographically incorrect form as long as they are phonologically accurate (e.g., *Hunt* for *Hund* [dog]). In contrast, the spelling book approach aims at establishing correct spellings from the beginning with easy, transparent words at first and by gradually and systematically increasing the complexity of the writing material. In the classes participating in Studies 1, 2, and 3, the spelling book approach was the only one used. In Study 4, no information about the instructional approach for teaching reading and spelling was collected but it seems safe to assume that the spelling book approach was the most prevalent one. Future studies could address the question

whether classes using different curricula than the spelling book approach differ in their error detection skills.

A general limitation of the present studies is that the data collection only took place in Bavaria – Lower Franconia (Studies 1, 2 and 3) and the greater Munich area (Study 4). Considering that the level of spelling skills can exhibit regional differences, the study samples are not representative for German students in general. However, there is no reason to assume that the relationships of error dictation and active spelling that are central for the present studies differs depending on the particular region where the data was collected.

Given that a computerized EDT allows automatically recording of response latencies, it is possible to analyze the efficiency of the error detection ability rather than simply the accuracy (Endlich et al., 2021). Further studies could examine this approach more systematically in order to clarify whether the efficiency might provide valuable information exceeding the accuracy in identifying misspelled words. Theoretically, it seems plausible to assume that high-quality, easily accessible lexical representations underlie both skilled spelling and reading (cf. the lexical quality hypothesis, Perfetti & Hart, 2002). For assessing reading skills, efficiency-based approaches that utilize both accuracy and reaction times have already proven their diagnostic utility (e.g., T. Richter, Isberner, Naumann, & Kutzner, 2012; T. Richter, Isberner, Naumann, Neeb, & Knoepke, 2017).

Finally, the error detection task used in the present studies is a screening instrument. Based on the two categories of misspelled words used in the present EDT – phonologically correct vs. incorrect misspellings – the next step would be to apply the method to a more detailed and comprehensive qualitative analysis of spelling errors as it is typically done with dictation and other tasks that involve the production of written texts. Error detection might provide a highly standardized and economical approach for teachers and learning therapists to diagnose spelling strategies and deficits in spelling development that should be addressed

with appropriate interventions, depending on the dominant type of spelling errors. Our results do not allow any conclusions regarding the use of error detection tasks for this purpose, but the strong convergence of error detection performance with dictation tasks provides promising evidence that the approach can also be used for assessing types of spelling errors and diagnosing spelling strategies.

Conclusion

The current results support the general theoretical assumption that skilled reading and spelling share a common knowledge base: Skilled readers and spellers rely on high-quality, accessible orthographic knowledge stored in the mental lexicon. Therefore, it is possible to assess spelling skills with an error detection task that does not involve writing but only reading, especially its component process of matching written words with orthographical word forms stored in the mental lexicon. An error detection task can be considered as an economical procedure for assessing spelling skills in German-speaking countries in elementary school. Assessing spelling skills with dictation tasks takes time for administering the test and evaluating the dictations, which is often associated with objectivity issues, whereas a tablet-based error detection task provides the students' with fully objective results as soon as the test has been finished. With the increasing availability of tablets in schools, teachers could use a tablet-based error detection task like the one introduced in the present study to obtain a quick overview of the spelling abilities of a whole class. Moreover, evaluation of the performances will then be completely objective, in contrast to correcting dictations (Birkel, 2009). The advantage of this economical procedure is even more apparent for the screening of children with poor spelling abilities.

References

Allen, D., & Ager, J. (1965). A factor analytic study of the ability to spell. *Educational and Psychological Measurement, 25*(1), 153–161.

Allred, R. A. (1984). Comparison of proofreading-type standardized spelling tests and written spelling test scores. *The Journal of Educational Research, 77*(5), 298–303. https://doi.org/10.1080/00220671.1984.10885544

Berninger, V. W., Abbott, R. D., Augsburger, A., & Garcia, N. (2009). Comparison of pen and keyboard transcription modes in children with and without learning disabilities. *Learning Disability Quarterly, 32*(3), 123–141. https://doi.org/10.2307/27740364

Birkel, P. (2009). Rechtschreibleistung im Diktat – eine objektiv beurteilbare Leistung? [Spelling performance in dictation – an objectively assessable performance?] *Didaktik Deutsch, 27*, 5–32.

Bruhn, K., Gudat-Vasak, S., Hinze, G., Müller, S., Nabers, B., & Reinker, D. (2014). *Bausteine – Fibel*. Braunschweig, Germany: Diesterweg.

Burt, J. S., & Tate, H. (2002). Does a reading lexicon provide orthographic representations for spelling? *Journal of Memory and Language, 46*(3), 518–543. https://doi.org/10.1006/jmla.2001.2818

Coltheart, M. (1978). Lexical access in simple reading tasks. In G. Underwood (Ed.), *Strategies of information processing* (pp. 151–216). San Diego, CA: Academic Press.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108*(1), 204–256. https://doi.org/10.1037/0033-295X.108.1.204

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

Endlich, D., Lenhard, W., Marx, P., & Richter, T. (2021). Tabletbasierter Fehleridentifikationstest zur ökonomischen und validen Erfassung von

Rechtschreibfähigkeiten in der Grundschule [Tablet-based error detection test as an economical and valid task for assessing spelling skills in elementary school]. *Lernen und Lernstörungen, 10*(1). https://doi.org/10.1024/2235-0977/a000324

Freyberg, P. S. (1970). The concurrent validity of two types of spelling tests. *British Journal of Educational Psychology, 40*(1), 68–71. https://doi.org/10.1111/j.2044-8279.1970.tb02100.x

Frisbie, D. A., & Cantor, N. K. (1995). The validity of scores from alternative methods of assessing spelling achievement. *Journal of Educational Measurement, 32*(1), 55–78. https://doi.org/10.1111/j.1745-3984.1995.tb00456.x

Grund, M., Haug, G., & Naumann, C. L. (1994). *Diagnostischer Rechtschreibtest für 4. Klassen (DRT 4)* [Diagnostic spelling test for Grade 4]. Weinheim, Germany: Beltz.

Guiler, W. S. (1929). Validation of methods of testing spelling. *Journal of Educational Research, 20*(3), 181–189. https://doi.org/10.1080/00220671.1929.10879981

Hassler Hallstedt, M., & Ghaderi, A. (2018). Tablets instead of paper-based tests for young children? Comparability between paper and tablet versions of the mathematical Heidelberger Rechen Test 1-4. *Educational Assessment, 23*(3), 195–210. https://doi.org/10.1080/10627197.2018.1488587

Holmes, V. M., & Babauta, M. L. (2005). Single or dual representations for reading and spelling? *Reading and Writing, 18*(3), 257–280. https://doi.org/10.1007/s11145-004-8129-5

Holmes, V. M., & Carruthers, J. (1998). The relation between reading and spelling in skilled adult readers. *Journal of Memory and Language, 39*(2), 264–289. https://doi.org/10.1006/jmla.1998.2583

Houghton, G., & Zorzi, M. (2003). Normal and impaired spelling in a connectionist dual-route architecture. *Cognitive Neuropsychology, 20*(2), 115–162. https://doi.org/10.1080/02643290242000871

Jiménez, J. E., Marco, I., Suárez, N., & González, D. (2017). Internal structure and development of keyboard skills in Spanish-speaking primary school children with and without LD in writing. *Journal of Learning Disabilities, 50*(5), 522–533. http://dx.doi.org/10.1177/0022219416633864

Jimerson, S., Egeland, B., Sroufe, L. A., & Carlson, B. (2000). A prospective longitudinal study of high school dropouts examining multiple predictors across development. *Journal of School Psychology, 38*(6), 525–549. http://dx.doi.org/10.1016/S0022-4405(00)00051-0

Klicpera, C., & Gasteiger-Klicpera, B. (1998). *Lesen und Schreiben. Entwicklung und Schwierigkeiten* [Reading and writing. Development and difficulties] (2nd ed.). Bern, Switzerland: Huber.

KMK (2005). *Beschlüsse der Kultusministerkonferenz. Bildungsstandards im Fach Deutsch für den Primarbereich. Beschluss vom 15.10.2004* [Decisions of the conference of ministers of culture. Educational standards in German for primary education. Resolution of October 15, 2004]. München, Germany: Wolters Kluwer.

Lenhart, J., Marx, P., Segerer, R., & Schneider, W. (2019). Rechtschreibung ohne Schreiben. Messen Fehleridentifikation und Diktat dasselbe? [Spelling without writing: Do error detection and dictation measure the same?] *Diagnostica*. Advance online publication. https://doi.org/10.1026/0012-1924/a000229

Loeber, R., & Dishion, T. J. (1983). Early predictors of male delinquency: A review. *Psychological Bulletin, 94*(1), 68–98. https://doi.org/10.1037/0033-2909.94.1.68

Mannhaupt, G. (1994). Deutschsprachige Studien zur Intervention bei Lese-Rechtschreibschwäche: Ein Überblick über neuere Forschungstrends [German-speaking studies on interventions in reading-spelling disorders: An overview of new research trends]. *Zeitschrift für Pädagogische Psychologie, 8*(3–4), 123–138.

May, P. (2012). *Hamburger Schreib-Probe 1-10 (HSP 1-10)* [Hamburg spelling probe 1-10] (6th ed.). Hamburg, Germany: Verlag für pädagogische Medien.

Müller, R. (1996). *Diagnostischer Rechtschreibtest für 3. Klassen (DRT 3)* [Diagnostic spelling test for Grade 3] (2th ed.). Göttingen, Germany: Beltz.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*(3), 412–433. https://doi.org/10.1037/met0000144

Metze, W. (2009). *Tobi. Erstlesebuch* [First reading book]. Berlin, Germany: Cornelsen.

Perfetti, C., & Hart, L. (2002). The lexical quality hypothesis. In L. Verhoeven, C. Elbro & P. Reitsma (Eds.), *Precursors of functional literacy* (pp. 189-213). Amsterdam, The Netherlands: John Benjamin.

Rapcsak, S. Z., Henry, M., Teague, S. L., Carnahan, S. D., & Beeson, P. M. (2007). Do dual-route models accurately predict reading and spelling performance in individuals with acquired alexia and agraphia? *Neuropsychologia, 45*(11), 2519–2524. https://doi.org/10.1016/j.neuropsychologia.2007.03.019

Reichen, J. (2006). *Hannah hat Kino im Kopf: Die Reichen-Methode "Lesen durch Schreiben" und ihre Hintergründe für LehrerInnen, Studierende und Eltern* [Hannah has cinema in her head: The Reichen method "reading by writing" and its background for teachers, students and parents]. Hamburg, Germany: Heinewetter.

Richter, S. (1994). Geschlechterunterschiede in der Rechtschreibentwicklung von Kindern der 1. bis 5. Klasse [Gender differences in the development of spelling skills in children from Grade 1 to Grade 5]. In S. Richter & H. Brügelmann (Eds.), *Mädchen*

lernen anders lernen Jungen. Geschlechtsspezifische Unterschiede beim Schriftspracherwerb (pp. 51–65). Bottinghofen, Germany: Libelle.

Richter, T., Isberner, M.-B., Naumann, J., & Kutzner, Y. (2012). Prozessbezogene Diagnostik von Lesefähigkeiten bei Grundschulkindern [Process-based assessment of reading skills in primary school children]. *Zeitschrift für Pädagogische Psychologie, 26*(4), 313-331. https://doi.org/10.1024/1010-0652/a000079

Richter, T., Lenhard, W., Marx, P., & Endlich, D. (2018). Konzeption eines Online-Screenings für Lernstörungen. *Lernen und Lernstörungen, 7*(4), 203-207. https://doi.org/10.1024/2235-0977/a000237

Richter, T., Naumann, J., Isberner, M.-B., Neeb, Y., & Knoepke, J. (2017). *ProDi-L: Prozessbasierte Diagnostik von Lesefähigkeiten bei Grundschulkindern* [Process-based assessment of reading skills in primary school children]. Göttingen, Germany: Hogrefe.

Romani, C., Olson, A., & Di Betta, A. M. (2005). Spelling disorders. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 431–447). Oxford, UK: Blackwell.

Schneider, M., Martinez Méndez, R., & Hasselhorn, M. (2014). *Fehleridentifikationstest: Rechtschreibung für fünfte und sechste Klassen (R-FIT 5-6+)* [Error detection test: Spelling for Grades 5 and 6]. Göttingen, Germany: Hogrefe.

Schneider, W. (1994). Geschlechtsunterschiede beim Schriftspracherwerb: Befunde aus den Müncher Längsschnittstudien LOGIK und SCHOLASTIK [Gender differences in literacy acquisition: Findings from the Munich longitudinal studies LOGIK and SCHOLASTIK]. In S. Richter & H. Brügelmann (Eds.), *Mädchen lernen anders lernen Jungen. Geschlechtsspezifische Unterschiede beim Schriftspracherwerb* (pp. 71–82). Bottinghofen am Bodensee, Germany: Libelle.

Schneider, W. (2008a). Entwicklung und Erfassung der Rechtschreibkompetenz im Jugend- und Erwachsenenalter [Development and assessment of spelling competence in youth and adulthood]. In W. Schneider, H. Marx & M. Hassehorn (Eds.), *Diagnostik von Rechtschreibleistungen und -kompetenz* (pp. 145–157). Göttingen, Germany: Hogrefe.

Schneider, W. (2008b). *Entwicklung von der Kindheit bis zum Erwachsenenalter: Befunde der Münchner Längsschnittstudie LOGIK* [Development from childhood to adulthood: Findings from the Munich longitudinal study LOGIK]. Weinheim, Germany: Beltz.

Schneider, W., & Näslund, J. C. (1999). The impact of early phonological processing skills on reading and spelling in school: Evidence from the Munich Longitudinal Study. In F. E. Weinert & W. Schneider (Eds.), *Individual development from 3 to 12: Findings from the Munich Longitudinal Study* (pp. 126–147). Cambridge: Cambridge University Press.

Schroeder, S., Würzner, K.-M., Heister, J., Geyken, A., & Kliegl, R. (2014). childLex: A lexical database of German read by children. *Behavior Research Methods, 47*(4), 1085–1094. https://doi.org/10.3758/s13428-014-0528-1

Staatsinstitut für Schulqualität und Bildungsforschung (2005). *Formen der Leistungserhebung im Fach Deutsch* [Forms of performance assessment in German]. Donauwörth, Germany: Auer.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*(2), 245–251. https://doi.org/10.1037/0033-2909.87.2.245

Stock, C., & Schneider, W. (2008a). *Deutscher Rechtschreibtest für das dritte und vierte Schuljahr (DERET 3-4+)* [German spelling test for third and fourth grade]. Göttingen, Germany: Hogrefe.

Stock, C., & Schneider, W. (2008b). *Deutscher Rechtschreibtest für das erste und zweite Schuljahr (DERET 1-2+)* [German spelling test for first and second grade]. Göttingen, Germany: Hogrefe.

Thomé, G., & Thomé, D. (2017). *Oldenburger Fehleranalyse für die Klassen 3–9 (OLFA 3–9). Instrument und Handbuch zur Ermittlung der orthografischen Kompetenz und Leistung aus freien Texten für die Entwicklung effektiver Fördermaßnahmen* [Oldenburger error analysis for Grades 3–9. Instrument and manual for the assessment of orthografic competence and performance from free writing for the development of effective supporting measures] (5th, impr. ed.). Oldenburg, Germany: Institut für sprachliche Bildung.

Westwood, P. (1999). The correlation between results from different types of spelling test and children's spelling ability when writing. *Australian Journal of Learning Disabilities, 4*(1), 31–36. https://doi.org/10.1080/19404159909546584

Author Note

This research was funded by the German Federal Ministry of Research (Bundesministerium für Bildung und Forschung, BMBF) in the project LONDI (Development of an Online Platform for Diagnosis and Remediation of Children with Learning Disabilities). The data and analysis scripts for all four studies are available https://osf.io/pqade/?view_only=e9e32a966ea24376a682d41f0b063221

Endlich, Lenhard, Marx, & Richter (2021) partially refer to the same data set.

Table 1

Overview of Studies

Study	Time	Instruments	Data collection	Participants
Study 1	July 2018; end of school year	EDT; dictation	Children's classroom	Grade 3: *n* = 45; Grade 4: *n* = 53
Study 2	December 2018; 3 months after beginning of school year	EDT; gap-fill dictation	Children's classroom	Grade 3: *n* = 52; Grade 4: *n* = 55
Study 3	January 2019; 4 months after beginning of school year	EDT; dictation; gap-fill dictation	Children's classroom	Grade 4: *n* = 54
Study 4	July 2020; end of school year	EDT; dictation	Online, at home	Grade 3: *n* = 646; Grade 4: *n* = 608

Note. EDT = Error detection task.

Table 2

Descriptive Statistics and Correlation Coefficients for Error Detection Task and Dictation at Grade 3 and 4 for Study 1, Study 2 and Study 4

		Grade 3								Grade 4							
	Test	*M*	*SD*	*Min*	*Max*	1	2	3	4	*M*	*SD*	*Min*	*Max*	1	2	3	4
Study 1	1 EDT-score (texts 1, 2 and 3)	28.82	14.71	4	60	-				16.25	10.07	2	47	-			
	2 Sum of identified misspelled words in EDT (% correct)	57.43 (68.4%)	14.63	25	82	-.99***	-			69.69 (81.1%)	9.91	39	84	-.99***	-		
	3 Marked correctly spelled words („false alarms“) in EDT	1.25	1.83	0	10	.10	.02	-		0.94	1.32	0	6	.19	-.06	-	
	4 Continuous dictation (misspelled words in DERET)	16.41	10.60	0	42	.71***	-.71***	.03		13.88	9.55	0	43	.76***	-.73***	.29*	
	5 Continuous dictation (DERET *T*-value)	54.12	11.21	31	73	-.72***	.72***	-.05	-.99***	53.60	9.57	31	73	-.76***	.73***	-.27	-.98***
Study 2	1 EDT-score (texts 1, 2 and 3)	48.85	13.65	20	74	-				34.13	14.87	10	67	-			
	2 Sum of identified misspelled words in EDT (% correct)	39.44 (46.9%)	12.60	13	67	-.97***	-			52.84 (62.9%)	14.35	25	77	-1.00***	-		
	3 Marked correctly spelled words („false alarms“) in EDT	2.29	3.29	0	18	.43**	-.20	-		0.96	1.30	0	6	.44**	-.36**	-	
	4 Gap-fill dictation (misspelled words in DERET)	10.19	4.36	2	19	.77***	-.77***	.25	-	12.96	6.04	2	24	.82***	-.82***	.36**	-
Study 4	1 EDT-score (texts 2 and 3)	13.77	9.04	0	48	-				9.72	7.33	0	41	-			
	2 Sum of identified misspelled words in EDT (% correct)	40.78 (75.5%)	8.94	6	54	-.99***	-			44.80 (83.0%)	7.20	14	54	-.99***	-		
	3 Marked correctly spelled words („false alarms“) in EDT	0.55	0.88	0	7	.16***	-.07	-		0.53	1.06	0	17	.20***	-.05	-	
	4 Continuous dictation (WRT *T*-value)	51.31	10.35	19	77	-.77***	.76***	-.16***	-	52.13	10.61	20	78	-.71***	.71***	-.11**	-

Note. EDT = Error detection task. WRT = Weingartener Grundwortschatz Rechtschreib-Test. DERET = Deutscher Rechtschreibtest. Different versions of the dictation tasks were used in Grade 3 and Grade 4. Study 1, Grade 3: $n = 44$. Study 1, Grade 4: $n = 51$. Study 2, Grade 3: $n = 52$. Study 2, Grade 4: $n = 55$. Study 4, Grade 3: $n = 646$. Study 4, Grade 4: $n = 605$. * $p < .05$, ** $p < .01$, *** $p < .001$.

Table 3

Error Detection as Predictor of Spelling Competencies Assessed with a Dictation Task (Study 1: DERET T-scores; Study 2: DERET z-score)

		Spelling competencies (Study 1: continuous dictation; Study 2: gap-fill dictation)			
		Grade 3		Grade 4	
	Variable	*B*	95% CI	*B*	95% CI
Study 1	Intercept	69.80**	[63.40, 76.21]	64.57**	[60.71, 68.43]
	EDT	-0.55**	[-0.72, -0.38]	-0.73**	[-0.91, -0.55]
	gender	0.34	[-4.57, 5.26]	1.46	[-2.14, 5.06]
	R^2	.526**		.581**	
	F	22.75		33.29	
Study 2	Intercept	-3.04**	[-3.84, -2.24]	-1.84**	[-2.34, -1.35]
	EDT	0.06**	[0.05, 0.07]	0.05**	[0.04, 0.07]
	gender	0.23	[-0.15, 0.61]	-0.05	[-0.38, 0.28]
	R^2	.611**		.679**	
	F	38.55		54.95	

Note. EDT = Error detection task. DERET = Deutscher Rechtschreibtest für das dritte und vierte Schuljahr; Stock & W. Schneider, 2008a.
B = unstandardized regression weights. CI = confidence interval. Study 1, Grade 3: $n = 44$. Study 1, Grade 4: $n = 51$. Study 2, Grade 3: $n = 52$. Study 2, Grade 4: $n = 55$.
* $p < .05$, ** $p < .01$.

Table 4

Error Detection as Predictor for Children with Poor Spelling Abilities Assessed with a Dictation Task (Study 1: Grades 3 and 4)

			Criterion variable: Dictation (DERET)	
			Poor spelling abilities (T-score < 40)	Uncritical spelling abilites (T-score ≥ 40)
Predictor variable: Error Detection Task	< 25th percentile		9[a]	20[b]
	≥ 25th percentile		1[c]	65[d]
Predictive values	Sensitivity	90.0%		
	Specificity	76.5%		
	Positive predictive value	31.0%		
	RIOC	85.6%		

Note. N = 95. DERET = Deutscher Rechtschreibtest für das dritte und vierte Schuljahr; Stock & W. Schneider, 2008a.
[a]true positive; [b]false positive; [c]false negative; [d]true negative.

Table 5

Error Detection and Gap-fill Dictation as Predictors of Spelling Competencies Assessed with a Continuous Dictation Task (Study 3)

	Spelling competencies (DERET)					
	Model 1		Model 2		Model 3	
Variable	*B*	95% CI	*B*	95% CI	*B*	95% CI
Intercept	7.86**	[5.29, 10.43]	3.65*	[0.79, 6.52]	3.46*	[0.75, 6.17]
Gap-fill dictation	2.50**	[1.76, 3.24]			1.10**	[0.28, 1.91]
EDT			0.48**	[0.37, 0.58]	0.35**	[0.22, 0.49]
R^2	.467**		.601**		.651**	
F	45.62		78.28		47.57	
ΔR^2	.184**		.050**			

Note. EDT = Error detection task. DERET = Deutscher Rechtschreibtest für das dritte und vierte Schuljahr; Stock & W. Schneider, 2008a.
B = unstandardized regression weights. CI = confidence interval. *N* = 54.
* $p < .05$, ** $p < .01$.

Table 6

Descriptive Statistics and Correlation Coefficients for Error Detection Task, Dictation and Gap-fill Dictation at Grade 4 (Study 3)

	Test	*M*	*SD*	*Min*	*Max*	1	2	3	4
Grade 4	1 EDT-score	23.74	11.96	6	58	-			
	2 Sum of identified misspelled words in EDT	61.31	12.21	26	78	-.99***	-		
	3 Marked correctly spelled words („false alarms“) in EDT	1.06	1.53	0	7	-.10	.22	-	
	4 Continuous dictation (misspelled words)	14.94	7.34	3	40	.78***	-.74***	.12	-
	5 Gap-fill dictation (misspelled words)	2.83	2.01	0	9	.67***	-.65***	.06	.68***

Note. EDT = Error detection task. $N = 54$.
*** $p < .001$.

Table 7

Error Detection as Predictor of Spelling Competencies Assessed with a Dictation Task (Study 4: WRT scores)

	Spelling competencies			
	Grade 3		Grade 4	
Variable	*B*	95% CI	*B*	95% CI
Intercept	63.09**	[62.00, 64.17]	62.26**	[61.08, 63.44]
EDT	-0.87**	[-0.93, -0.82]	-1.02	[-1.10, -0.94]
gender	0.56	[-0.47, 1.59]	-0.23	[-1.40, 0.95]
R^2	.588**		.509**	
F	459.5		311.8	

Note. EDT = Error detection task. DERET = Deutscher Rechtschreibtest für das dritte und vierte Schuljahr; Stock & W. Schneider, 2008a.
B = unstandardized regression weights. CI = confidence interval. Grade 3: $n = 646$. Grade 4: $n = 605$.
* $p < .05$, ** $p < .01$.

Table 8

Correlation Coefficients for Error Detection Task, Dictation and Reading Fluency at Grade 3 and 4 (Study 4)

	Test	*M*	*SD*	1	2	3
Grade 3	1 EDT-score	13.77	9.04	-		
	2 Dictation (*T*-value)	51.31	10.35	-.77***	-	
	3 Reading fluency (*T*-value)	55.80	8.79	-.47***	.56***	-
Grade 4	1 EDT-score	9.72	7.33	-		
	2 Dictation (*T*-value)	52.13	10.61	-.71***	-	
	3 Reading fluency (*T*-value)	53.96	8.31	-.51***	.57***	-

Note. EDT = Error detection task. Grade 3: $n = 646$. Grade 4: $n = 605$.
*** $p < .001$.

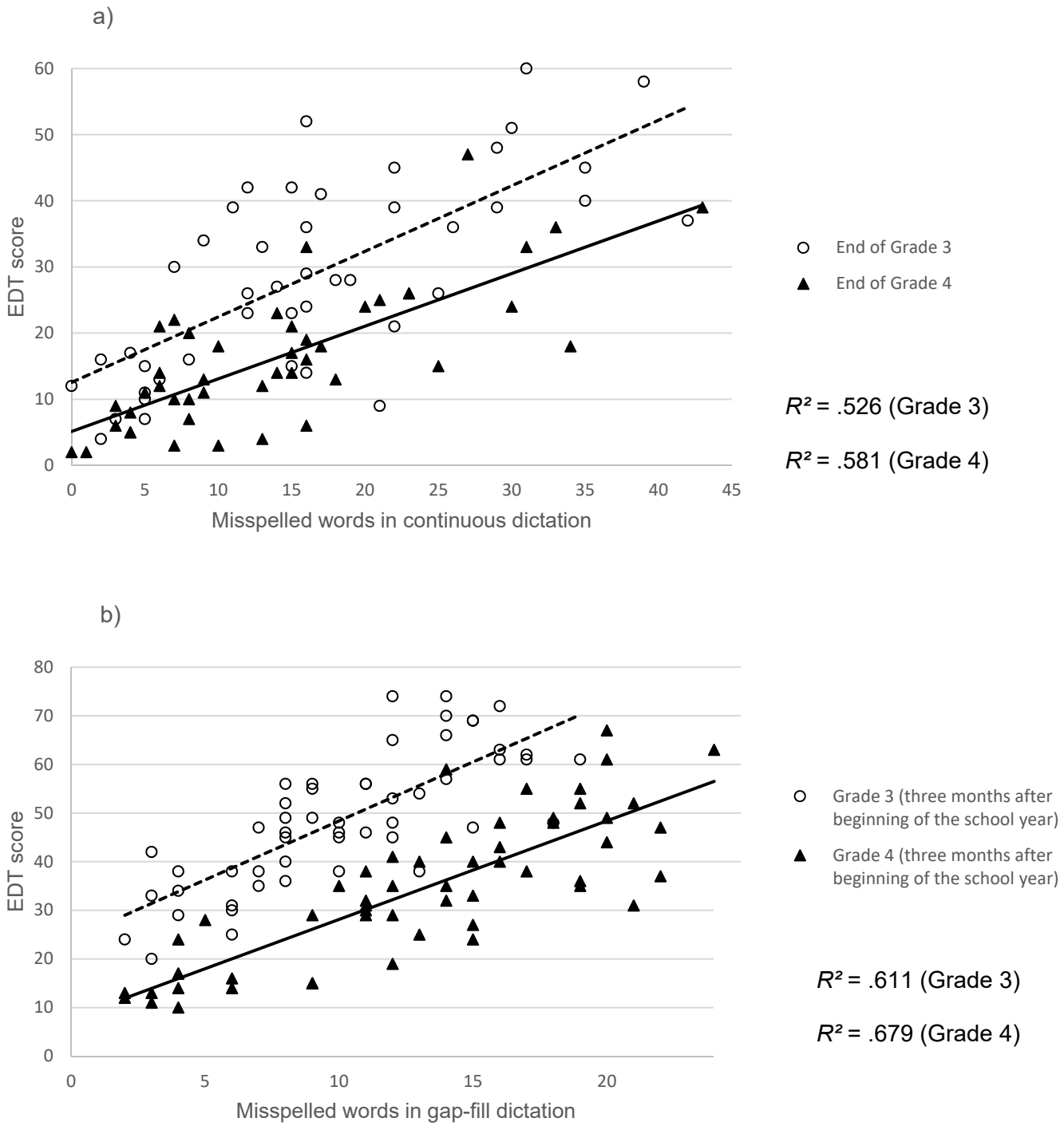

Figure 1. Scatter plots showing the association between performance in the error detection task (EDT score) and different types of dictation tasks (continuous dictation in Study 1 (a) and gap-fill dictation in Study 2 (b)) at different times in Grade 3 and Grade 4.

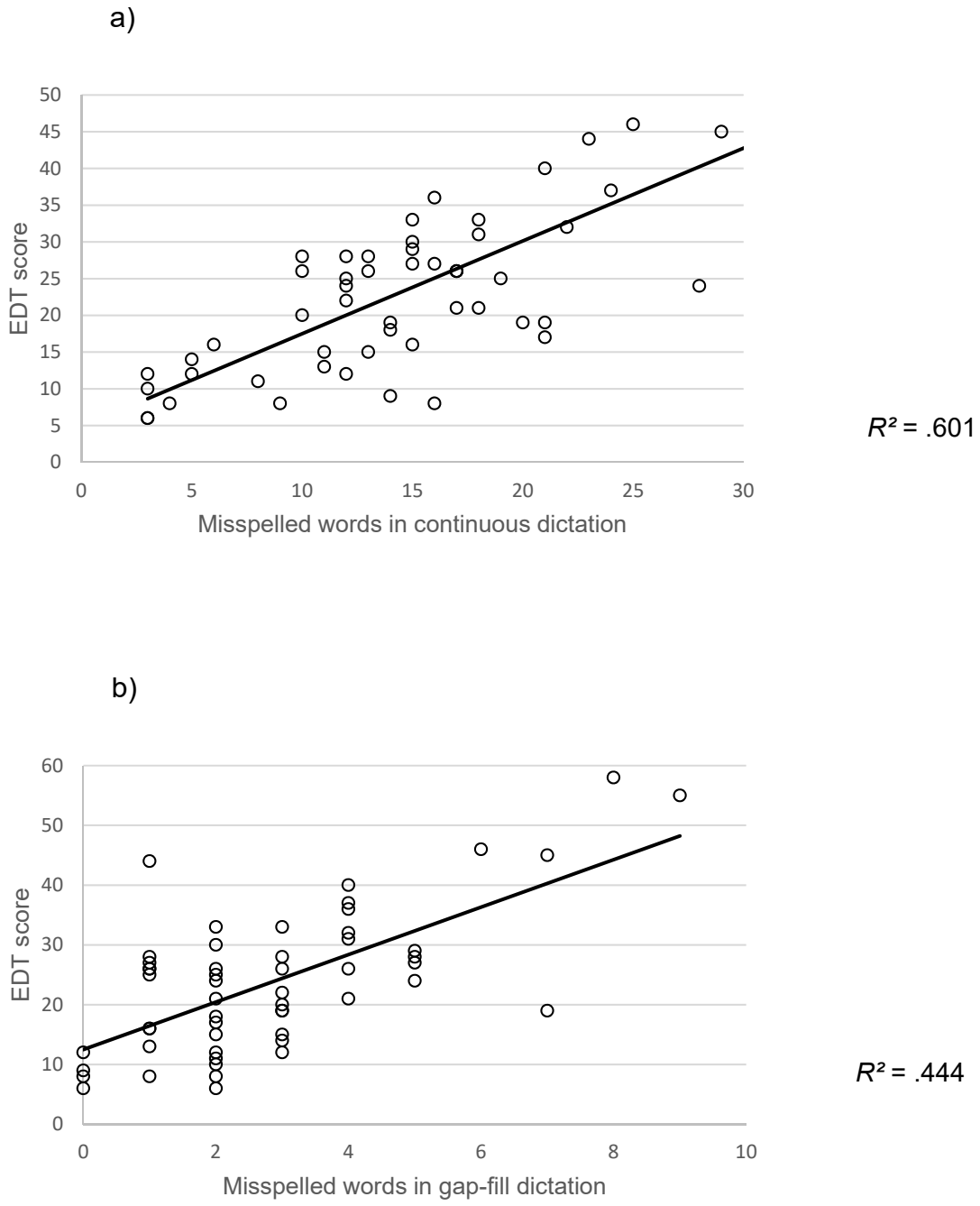

Figure 2. Scatter plots showing the association between performance in the error detection task (EDT score) and different types of dictation tasks (continuous dictation (a) and gap-fill dictation (b)) in Grade 4 (four months after beginning of the school year; Study 3).