Instability of word reading errors of typical and poor readers

Esther G. Steenbeek-Planting, Wim H. J. van Bon and Robert Schreuder†
Radboud University Nijmegen

We examined the instability of reading errors, that is whether a child reads the same word sometimes correctly and sometimes incorrectly, and whether typical readers differ in their instability from poor readers. With an interval of a few days, Dutch CVC words were read twice by typically developing first and second graders and reading-level matched poor readers. Error instability was considerable and second graders produced more unstable errors than first graders. Poor readers did not differ from typical readers, suggesting a developmental lag for poor readers. Of the word characteristics studied, frequency was the strongest predictor: the higher word frequency, the higher error instability. Our study indicates that error instability can be considered as an indicator of the transition from incompetence to reading competence.

Keywords: word decoding, reading development, reading disorders, orthography, reading errors, error instability

1. Introduction

The misreading of individual words by readers of low-level skill is often taken to indicate either ignorance about conversion rules or the lack of word specific knowledge, necessitating instruction or practice on that specific rule or on the peculiarities of the individual words. The conclusion, however, that a misreading points to a lack of general or specific knowledge clearly is not warranted if a misreading is only accidental and was not made at an earlier occasion or is not repeated when reading the word anew. The analysis of children's reading errors can help us understand how they learn to read (Goikoetxea, 2006). Practitioners in education settings are encouraged to analyze the reading errors (miscues) made in order to detect possible patterns of errors, which may provide a window on

Written Language & Literacy 20:2 (2017), **147–169**. DOI 10.1075/wll.00002.ste
ISSN 1387–6732 / E-ISSN 1570–6001

inefficient reading strategies the child applies. The error patterns thereby may guide the instruction and remediation of poor readers in particular (McKenna & Picard, 2006). Error patterns are thus assumed to stem from a specific lack of competence and, as a rule, are not assumed to stem from mere inattention on the part of the child. Theoretically, reading error instability would indicate that reading is only partly determined by stable knowledge, and is prone to factors of a rather accidental nature that deserve a place in models for reading performance. If this 'unstable pattern' marks the transition from incompetence to competence it should be a sensitive index for factors in reading development. The instability of reading errors, that is, how often words are read incorrectly on one occasion and correctly on another, is explored.

Although a prime characteristic of reading disorders in languages with a transparent orthography is the impairment in reading speed (de Jong & van der Leij, 2003; Landerl, 2001; Serrano & Defior, 2008), accuracy is also affected (Patel, Snowling, & de Jong, 2004). Error patterns of beginning readers have shown to reflect the linguistic and orthographic system (Cossu, Shankweiler, Liberman, & Gugliotta, 1995; Ellis et al., 2004; Ognjenović, Lukatela, Feldman, & Turvey, 1983;), revealing that alphabetic processing is a basic reading strategy in readers of a transparent orthography (Aro & Wimmer, 2003; Goswami, 2002; Ellis et al., 2004; Guron & Lundberg, 2004; Seymour, Aro, & Erskine, 2003; Wimmer & Goswami, 1994). Focus of the present paper therefore is on words with a phonological consonant-vowel-consonant (CVC) structure, because, these words – with the exception of a few loan words – are orthographically fully transparent in reading, that is, can be read by applying grapheme-phoneme correspondence rules (see Booij (1995) for an introduction to Dutch orthography).

Next to the authors (Steenbeek-Planting, van Bon, & Schreuder, 2013a), only Gough, Juel, and Griffith (1992) have directly examined the instability of reading errors made by typically developing children. They asked beginning English readers to read and spell the same words, which were mostly regular CVC words, in two sessions spread across a week. The children's reading errors were not found to be completely stable as only 70% of the words were read correctly on both occasions, 12% were read incorrectly on both occasions (stable errors), and 18% were read incorrectly on one occasion and correctly on another (unstable errors). It is uncertain whether Gough et al.'s findings for the opaque orthography of English, apply to the transparent orthography of Dutch as well. Our study fits in with Gough et al.'s study as we will use the same operationalization to define stable (twice incorrect) and unstable (once correct, once incorrect) errors. Participants in the present study are Dutch typically developing first- and second grade readers and reading-level matched poor readers.

Error instability might be affected by word characteristics such as frequency, bigram frequency, number and frequency of orthographic neighbors. First of all, word frequency strongly affects lexicalization (e.g., Murray & Forster, 2004; Rastle, 2007). Facilitative effects of sublexical word characteristics for adult word recognition such as bigram frequency have been reported by, among others, Arduino & Burani, (2004), Balota, Yap, & Cortese (2006), and Gernsbacher (1984). Bigram frequency correlates with neighborhood characteristics; words with a high bigram frequency usually not only have many neighbors, but also more high-frequency neighbors (Frauenfelder, Baayen, Hellwig, & Schreuder, 1993; Landauer & Streeter, 1973). A neighbor refers to a word that can be formed from another word by changing one letter. In our study, we will regard the number and the word frequency of neighbors.

In the present study, the main questions are: How often are words read unstably (i.e., how often is the same word read incorrectly on one occasion and correctly on another)? Which word characteristics determine the occurrence of reading errors and the instability of reading errors? Do typically developing readers in grade 1 and 2 differ in their instability from reading-level matched poor readers?

2. Method

2.1 Participants

Typical readers and reading-level matched poor readers were selected from seven schools: five regular primary schools and two special schools for primary education (Eurybase, 2008). All children had Dutch as their first language. Neither the manner in which the schools were selected nor incidental information on the individual schools suggested that the samples were biased by their socioeconomic background. Two versions of a lexical decision test (LDT, Van Bon, 2007) were employed to initially determine the reading abilities of both the students in regular and special schools for use in sample selection. We incorporated the LDT, as it is an adequate and reliable alternative for using oral reading tests (van Bon, Hoevenaars, & Jongeneelen, 2004). In the regular schools, all students in grades 1, 2, and 3[1] ($n = 673$) were asked to complete the LDT. In the special schools, all grade 1 through 6 students ($n = 386$) were asked to complete the LDT.

---

[1] Children in grade 3 were tested and assigned to either the group of poor readers that were reading-level matched to typical readers in grade 2 (PR2), or the group of age-matched typical readers (post-hoc analysis in the discussion).

Four reader groups were formed on the basis of the LDT: typical readers in grade 1 (TR1, $n = 47$); typical readers in grade 2 (TR2, $n = 46$); poor readers matched to the reading level of the typical grade 1 readers (PR1, $n = 40$); and poor readers matched to the reading level of the typical grade 2 readers (PR2, $n = 45$).

The LDT scores for 50 randomly-selected grade 1 students, 50 randomly-selected grade 2 students, and 50 randomly-selected grade 3 students in regular education were taken as a point of departure. The students with a score above the 10th percentile for their grade level on the LDT were considered typical readers; the remaining students were considered poor readers. Three poor readers in grade 2 were assigned to PR1, and four poor readers in grade 3 were assigned to PR2, because their reading scores were in the range of TR1 or TR2 respectively.

Next, poor readers in the special schools were selected for participation. Children are in this type of education because of their learning disabilities, mild mental retardation or mild behavioral problems, as autism, ADD or ADHD. The majority of these children (73%) are poor readers (van Bon, Bouwmans, & Broeders, 2006). Children with mental retardation or behavioral problems (e.g., articulatory, oculomotor, visual or hearing) were excluded from participation. Children were considered poor readers if they scored below the 10th percentile for their grade level on the LDT and if they had received at least one more year of reading instruction than the typical readers they were matched to. Of the special school students, 37 poor readers in grade 2, 3 and 4 were allocated to PR1, and 41 poor readers in grade 3, 4 and 5 were allocated to PR2, if their reading scores were in the range of TR1 or TR2 respectively.

The descriptive statistics for the four reader groups can be found in Table 1. As can be seen, the number of boys and girls was almost equal among the typical readers. The poor readers included more boys than girls (Habib, 2000).

To verify whether PR1 matched TR1, and PR2 matched TR2, selected participants completed a word decoding test (WDT) and a nonword reading test. A multivariate analysis of variance, with the reading scores of PR1 and TR1 on the LDT1 and 2, WDT1, 2 and 3, and nonword reading test as the dependent variables, and Reading group (typical vs. poor) as factor (see Table 1), showed a main effect of Reading group ($F(6, 58) = 2.53$, $p < .05$, $\eta_p^2 = .21$).[2] However, closer examination of the main effect showed PR1 to *not* differ from TR1 on any test (LDT1 and WDT1, 2 and 3 $F < 1$; LDT2: $F(1, 63) = 1.15$, $p = .29$, $\eta_p^2 = .02$; nonword reading test: $F(1, 63) = 1.93$, $p = .17$, $\eta_p^2 = .03$). It can thus be tentatively concluded that the reading performance on real words and nonwords of PR1 matched that of

---

[2] Due to listwise deletion in case of missing values, in the multivariate ANOVA some cases were missing. 65 Children participated in the first analysis (PR1 and TR1), and 78 children participated in the second analysis (PR2 and TR2).

Table 1. Descriptive statistics for four groups of readers; means and standard deviations (in parentheses) for paper-and-pen lexical decision tests with monosyllabic words (LDT1) and bisyllabic words (LDT2), word decoding test (WDT1, WDT2, and WDT3), and nonword reading (NWR). Scores are reported as items per minute

	Reader groups							
	TR1 (*n* = 47)		**PR1 (*n* = 40)**		**TR2 (*n* = 46)**		**PR2 (n = 45)**	
Gender (girls/boys)	**24/23**		**14/26**		**21/25**		**13/32**	
Age range^{a}	6;4 -	8;8	8;0 -	11;5	7;6 -	9;6	8;3 -	11;4
Age^{a}	7;2	(0;5)	9;5	(0;10)	8;1	(0;6)	10;3	(0;10)
Months of reading instruction	9	(0)	25	(7)	19	(0)	32	(6)
Reading tests								
LDT1	22.55	(10.98)	21.75	(7.87)	40.96	(11.38)	40.63	(11.44)
LDT2	18.29	(12.39)	21.45	(9.00)	43.15	(19.29)	47.49	(17.38)
WDT1	48.59	(23.11)	42.25	(19.57)	76.46	(13.78)	75.58	(18.23)
WDT2	33.11	(22.75)	27.58	(16.87)	66.11	(19.00)	63.29	(21.61)
WDT3	19.68	(16.07)	17.33	(13.75)	48.34	(16.71)	47.76	(19.78)
NWR	12.13	(6.96)	9.22	(5.27)	21.24	(6.42)	18.01	(7.85)

Note. Age is given in years; months. TR1 = Typical Readers in Grade 1, TR2 = Typical Readers in Grade 2, PR1 = Poor readers matched to the reading level of TR1, PR2 = Poor Readers matched to the reading level of TR2.

TR1. Next, a similar analysis was performed to verify whether PR2 matched TR2. Again, we found a main effect of Reading Group (F (6,67) = 2.32, $p < .05$, $\eta_p^2 = .17$, but PR2 did not differ from TR2 on any of the tests (LDT1 and 2, WDT1, 2 and 3 $F < 1$; nonword reading: F (1, 72) = 3.15, $p = .08$, $\eta_p^2 = .04$). It can thus be tentatively concluded that PR2's reading performance matched TR2 as well.

2.2 Procedure and instruments

2.2.1 *Reading tests*

Lexical decision test (LDT). The students were asked to complete two versions of a standardized paper-and-pen lexical decision test (Van Bon, 2007). Each version involves a card with words distributed across it in columns. LDT1 is composed of CVCC and CCVC words and has 60 nouns interspersed with 20 pseudowords. LDT2 is composed of bisyllabic words and has 90 nouns interspersed with 30 pseudowords. Students are asked to silently read the items and cross out every pseudoword. The raw score for each test is the number of words judged within a minute minus the number of errors. The tests were administered in class by the teacher. Test – retest reliability for children in grades 1 to 3 is considered sufficient, .81 for LDT1 and .82 for LDT2, (Van Bon, 2007).

Word decoding test (WDT). A standardized word reading test, the 'Drie-Minuten-Toets' (Verhoeven, 1995) [Three One-Minute Tests] was administered individually to assess the oral reading abilities of the students for words in isolation. This test consists of three cards with words listed in columns (WDT1: simple monosyllabic words; WDT2: monosyllabic words with one or two consonant clusters; WDT3: two-, three-, and four-syllable words). Students are instructed to read the words aloud as quickly and accurately as possible. The raw score for each card is the number of words read correctly in one minute. The reported reliability of the three cards (Cronbach's α) ranges from .86 to .94 (Verhoeven & van Leeuwe, 2003) and is judged sufficient.

Nonword reading test. In order to assess the decoding ability of the students for pseudowords, a standardized nonword reading test was administered (van den Bos, Lutje Spelberg, Scheepstra, & de Vries, 1994). The test consists of nonwords of increasing length. The students are instructed to read the pseudowords aloud as quickly and accurately as possible. The test score is reported as the number of nonwords read correctly per minute. The parallel reliability is good, .93 and above (van den Bos et al., 1994).

2.2.2 *Computer reading task*

Stimuli for the computer reading task. Words with a phonological CVC structure were selected from the Celex Database (for a description see Baayen, Piepenbrock, & van Rijn (1993)). The selected words were the lemmas for words that can occur independently in a language (i.e., nouns, verbs, adjectives, adverbs, and numbers). Proper names and words with a foreign orthography or phonology were eliminated. Of the initial 1078 lemmas, 861 constituted the final set.[3]

Sampling from the lemma set. To ensure a sample that reflects the reading of words children generally encounter in print, the CVC words were sampled on the basis of their token count.[4] For all lemmas frequency was calculated using the Celex Database (Baayen et al., 1993). If necessary frequencies for the different syntactic classes of the same lemma were summed. Sampling chance of a lemma selection was proportional to its token frequency. Thus, the higher the frequency of occurrence of a lemma, the higher its possibility of being selected. Thirty different samples of 200 lemmas each were randomly drawn for use with individual participants.

2.2.3 *Procedure*

General procedure. All children were tested in the same period of the school year. The WDT and the computer reading task (first time) were administered on the first day of testing. The nonword reading test and the computer reading task (second time) were administered one to three days later.

Administration of the computer reading task. Laptops with 14" screens were used for the computer reading task. Each CVC target word was presented in black lower case letters (Arial, size 46) on a white background in the center of the screen. The letters strings had a height of approximately 1.5 cm and ranged from 2 cm to 6.5 cm in length. The child was seated approximately 60–80 cm from the computer screen. A microphone was positioned in front of the child.

The task was administered twice (Time 1 and Time 2). The same word sample was used on both occasions but the words were presented in a different random order. Each testing occasion started with a practice block of 20 randomly presented CV and VC words. Next, the 200 target words were presented in five blocks

---

[3] Some words were eliminated for more than one reason.

[4] A type represents a unique linguistic entry; a token represents every occurrence of a given type. A token list thus involves selection from a group of items representing various linguistic types proportional to their frequency of use in the language. A high frequency word will thus occur repeatedly in the set to be selected from and a low frequency word less or not repeatedly.

of 40 words. Each block was followed by a short break. The experimenter recorded the (in)correctness of the students' responses using a button box connected to the computer. All responses that did not correspond to the correct pronunciation of the word, were considered errors. If students corrected themselves, the response was not considered an error.

The target words were presented one at a time, and the child was instructed to name the word on the screen as quickly and as accurately as possible. Each item was preceded by the presentation of a fixation cross (a +) in the center of the screen for 750 ms. After a blank screen for 150 ms, the target word was presented. The word disappeared as soon as the student spoke. If the student did not speak within 10 seconds, the response was considered incorrect and exposure was terminated. No feedback was given.

Trials were considered invalid if the voice key was triggered by another sound than the student's voice or if the voice key did not respond. The percentage of invalid trials was 8.5% for PR1, 4.0% for PR2, 6.5% for TR1, and 4.4% for TR2.

2.3 Data analyses

Test score stability was based on the accuracy scores for each subject summed across items (i.e., numbers correct) at each occasion. Stability of item difficulty was based on the item scores averaged across subjects (Nunnally & Bernstein, 1994) at each occasion. The stability of both over the two test occasions were determined using the Intraclass Correlation (ICC, Absolute Agreement, Two-Way Mixed, Single Measure).

The amount of error instability is – almost inevitably – related to the number of errors: In case of only errors or of no errors at all, there can be no instability. Maximum instability can be obtained if a student's score or an item's difficulty is 50%. Therefore, we calculated an error instability measure that corrects the number of instabilities found for a student or an item for the theoretical maximum for this number of errors, the Instability Score (IS)[5] (Appendix A). IS, which can be calculated for subjects (IS_{subj}), and items (IS_{item}), can vary from 0 to 1. The score 0 will be interpreted as *maximally stable* and 1 as *maximally unstable*.

An IS could not be calculated for subjects who (a) made no errors (TR1: $n = 3$; TR2: $n = 7$; PR2: $n = 6$), or (b) made errors on *only* one occasion. Subjects without an IS for the latter reason were therefore classified as maximally unstable (TR1: $n = 4$; TR2: $n = 13$; PR1: $n = 1$; PR2: $n = 4$). Similarly, items were classified as maximally unstable that were misidentified at one occasion only (TR1: $n = 185$; TR2: $n = 118$; PR1: $n = 161$; PR2: $n = 158$).

---

[5] A kappa-like measure was not used to determine instability as such an outcome is biased by the number of errors as well.

3. Results

First, the stability of test scores and item difficulty is compared between groups of readers. Next, error instability is explored, for subjects and items separately.

3.1 Stability of reading scores

3.1.1 *Test score stability*

The percentages correct for the two administrations of the computer reading task were calculated. In the upper part of Table 2, the mean percentages correct at Time 1 and Time 2 are presented. The ICC between Time 1 and Time 2 were highly significant for all of the groups ($p < .001$), indicating that the accuracy of the students' reading was very stable.

Possible differences in the stability of the reading levels were examined using a repeated measures analysis of variance. Percentage correct was the dependent variable. Time (first vs. second administration) was a within-subjects factor, and Reading level (1 vs. 2) and Reading group (typical vs. poor) were between-subjects factors. The results showed main effects of both Reading level (F (1, 174) = 59.91,

Table 2. Mean reading accuracy and item difficulty per reading group on the computer reading task (SD in parentheses), results of the paired t-tests for Time 1 vs. Time 2, and Intraclass Correlations (ICC) between Time 1 and Time 2

Reading group	Mean percentage correct				*t*	*df*	ICC
	Time 1		Time 2				
			Reading accuracy (averaged across items)				
TR1	92.21	(8.37)	93.04	(7.90)	−1.54	46	$.90^{***}$
TR2	98.36	(1.78)	98.51	(2.04)	−.73	45	$.76^{***}$
PR1	87.22	(9.95)	86.88	(10.75)	.50	39	$.92^{***}$
PR2	96.73	(4.02)	96.70	(3.97)	.13	44	$.92^{***}$
			Item difficulty (averaged across subjects)				
TR1	91.56	(15.31)	92.63	(14.23)	−1.410	622	$.18^{**}$
TR2	97.43	(9.32)	98.08	(6.80)	−1.465	616	$.11^{*}$
PR1	87.11	(17.67)	86.93	(18.25)	.201	583	$.35^{**}$
PR 2	96.16	(10.35)	95.86	(10.88)	.687	614	$.48^{**}$

Note. TR1 = Typical Readers in Grade 1, TR2 = Typical Readers in Grade 2, PR1 = Poor readers matched to the reading level of TR1, PR2 = Poor Readers matched to the reading level of TR2.

* $p < .05$

** $p < .01$

*** $p < .001$.

$p < .001$, $\eta_p^2 = .26$) and Reading group ($F(1, 174) = 13.31$, $p < .001$, $\eta_p^2 = .07$), but no main effect of Time ($F < 1$, $\eta_p^2 < .01$), no interaction of Time by Reading level ($F < 1$, $\eta_p^2 < .01$), and no interaction of Time by Reading group ($F(1, 174) = 2.25$, $p = .14$, $\eta_p^2 = .01$). A trend towards significance was observed for the Reading level by Reading group interaction, however ($F(1, 174) = 3.72$, $p = .06$, $\eta_p^2 = .02$). No third order interaction of Time by Reading level by Reading group was found ($F(1, 174) = 1.23$, $p = .27$, $\eta_p^2 < .01$). The absence of effects involving Time attests again to the stability of the test scores.

3.1.2 *Stability of item difficulty*

Statistics on item difficulty – mean percentage of students who read a specific item correctly on Time 1 or Time 2 – are presented in the bottom of Table 2.[6] A repeated measures analysis of variance was conducted on the percentage correct per item (i.e., with items as cases), with Time (first vs. second administration) as a within-items factor, and Reading level (1 vs. 2) and Reading group (typical vs. poor) as between-subjects factors. The results showed main effects of both Reading level ($F(1, 2435) = 61.69$, $p < .001$, $\eta_p^2 = .11$) and Reading group ($F(1, 2435) = 284.24$, $p < .001$, $\eta_p^2 = .03$) but no main effect of Time ($F < 1$, $\eta_p^2 < .01$), no interaction of Time by Reading level ($F < 1$, $\eta_p^2 < .01$), and no interaction of Time by Reading group ($F(1, 2435) = 2.88$, $p = .09$, $\eta_p^2 < .01$). A significant Reading level by Reading group interaction was found ($F(1, 2435) = 14.64$, $p < .001$, $\eta_p^2 < .01$). No third order interaction was found of Time by Reading level by Reading group ($F < 1$, $\eta_p^2 < .01$). The absence of effects involving Time shows the accuracy levels for the items to not change systematically for any of the reader groups.

The ICCs between Time 1 and Time 2 were found to be significant for all of the groups but fairly low – particularly when compared to the ICCs for the test scores. The rather low stability suggests that the probability of a specific word being read correctly varies across occasions.

3.2 Instability of reading errors

3.2.1 *Subject analyses*

Tables 3a through 3d present the percentages of words read correctly or incorrectly on both occasions, and unstably.

---

[6] The mean item difficulty may differ slightly from the mean reading accuracy due to a few missing or invalid items.

Tables 3a–3d. Percentages of words read correctly and incorrectly at times 1 and 2 for four reading groups

Table 3a. Typical readers grade 1

	Time 2	
Time 1	**Incorrect**	**Correct**
Incorrect	2.41%	5.38%
Correct	4.56%	87.65%

Table 3b. Typical readers grade 2

	Time 2	
Time 1	**Incorrect**	**Correct**
Incorrect	0.34%	1.30%
Correct	1.15%	97.21%

Table 3c. Poor readers matched to grade 1

	Time 2	
Time 1	**Incorrect**	**Correct**
Incorrect	5.17%	7.61%
Correct	7.95%	79.26%

Table 3d. Poor readers matched to grade 2

	Time 2	
Time 1	**Incorrect**	**Correct**
Incorrect	0.95%	2.32%
Correct	2.35%	94.38%

The number of unstable responses varies from as low as 2.45% for TR2, to 15.56% for PR1. While the poor readers made more unstable errors on average than the typical readers, the percentage of unstable errors, as a function of the errors made, is larger for the typical readers than the matched poor readers, and larger in grade 2 reading level than in grade 1 reading level. It thus seems that the number of unstable errors need to be interpreted with regard to the total number of errors made. Both the number of stable and unstable errors are significantly predicted by the error percentages reported in Table 3 (as for stable errors: $r = .88$ for TR1, .74 for TR2, .90 for PR1 and .92 for PR2; as for unstable errors $r = .98$ for TR1, TR2, .and PR1, and .99 for PR2; $p < .001$ for all analyses).

In Table 4, the mean IS_{subj}, the *SD*s, and the results of one-sample t-tests to determine whether the IS_{subj} significantly deviates from maximally stable (0) or maximally unstable (1) are presented. For each group, the IS_{subj} was found to deviate significantly from both 0 and 1, which shows that the reading errors were neither completely stable nor completely unstable.

Table 4. Means and SDs for instability scores summed across subjects (IS_{item}) or across Items (IS_{subj}) and results of one-sample t-tests for IS_{subj} against 0 and 1, respectively

	IS_{subj}					IS_{item}		
				One-sample t-tests				
				(test value = 0)	**(test value = 1)**			
Reading group	*M*	*SD*	*df*	*t*	*t*	n^a	*M*	*SD*
TR1	.65	.29	43	14.63***	−7.91***	378 (623)	.75	.41
TR2	.82	.28	38	18.21***	−3.83***	162 (617)	.85	.35
PR1	.59	.23	39	16.37***	−11.16***	417 (584)	.68	.42
PR 2	.80	.21	38	23.59***	−5.82***	241 (615)	.81	.38

Note. TR1 = Typical Readers in Grade 1, TR2 = Typical Readers in Grade 2, PR1 = Poor readers matched to the reading level of TR1, PR2 = Poor Readers matched to the reading level of TR2.
a *n* denotes the number of items incorporated in the study, that is the number of items with valid Instability Scores.
* $p < .05$
** $p < .01$
*** $p < .001$.

An analysis of variance with IS_{subj} as the dependent variable and Reading level (1 vs. 2) and Reading group (typical vs. poor) as the between-subjects factors rendered no significant interaction of Reading level by Reading group ($F < 1$, $\eta_p^2 < .01$) and no main effect of Reading group ($F < 1$, $\eta_p^2 = .01$). Evidently, the instability of the reading errors produced by typical readers does *not* differ from the instability of the reading errors produced by reading-level matched poor readers. The main effect of Reading level ($F(1, 158) = 22.38$, $p < .001$, $\eta_p^2 = .12$) shows the reading errors of readers at level 2 to be *less* stable than the reading errors of readers at level 1.

In sum, the reading errors were not completely unstable, nor completely stable. Typical readers did not differ from reading-level matched poor readers. The reading errors of the readers at level 2, however, were less stable than those of readers at level 1.

3.2.2 *Item analyses*

The means and *SD*s for IS_{item} are presented in Table 4. Analyses were conducted to determine which word characteristics affect the occurrence and the instability of reading errors. The characteristics of interest were determined for each item (see Table 5): (1) word frequency (the natural logarithm of its frequency per million),

(2) bigram frequency, (3) number of orthographic neighbors, and (6) word frequency for the most frequent orthographic neighbor (the natural logarithm).

Given considerable collinearity in the predictor variables (the condition number was too high (69) according to Belsley, 1991), entry of all the variables into the regression analyses would affect the estimates of the coefficients and their variances in the linear models (Chatterjee, Hadi, & Price, 2000). To avoid collinearity, we regressed bigram frequency, neighborhood size, and frequency of the most frequent neighbor on word frequency. The original variables were replaced by their residualised ones (henceforth Bigram residuals, Neighborhood size residuals and Neighbor frequency residuals). These residuals are now uncorrelated with word frequency, but still highly correlated with their original raw scores. No collinearity was found among the replaced predictors: The condition index was low (2), and tolerance and the variance inflation factor were acceptable (respectively > .40 and < 2.5; see Allison, 1999).

Table 5. Descriptives of word characteristics (861 words)

Word characteristic	Min	Max	M	(SD)
Word frequency	−1.62	4.32	0.77	1.01
Bigram frequency	11.53	15.79	13.72	(0.67)
Frequency orthographic neighbor	−0.13	4.32	2.30	(0.74)
Number of orthographic neighbors	0	28	11.34	(4.69)

Hierarchical linear regression analyses were conducted for each reading group separately, with accuracy as the dependent variable. The predictor variables were Word frequency, Bigram residuals, Neighborhood size residuals, and Neighbor frequency residuals (see Table 6). Accuracy was found to be explained to only a limited degree (adjusted R^2 varying from .02 to .05), probably due to the high item scores. Word frequency and Neighborhood size residuals were the strongest predictors. Neighborhood frequency residuals was a significant predictor for PR2 only.

Linear regression analyses to determine which word characteristics affect the instability of reading errors were not permitted as the IS_{item} data were bimodally distributed with peaks towards stability (0) and instability (1). Therefore, IS_{item} was dichotomized into 0 for items that were read stably incorrect and 1 for all items that were read unstably, independent of the degree of instability. Binary logistic regression analyses for each of the reading groups with dichotomized IS_{item} as the dependent variable and Word frequency, Bigram residuals, Neighborhood size residuals, and Neighbor frequency residuals as the explanatory variables (see Table 6) revealed significant models for TR1, PR2, and marginally for PR1, but not for TR2. Using R^2 of McKelvey and Zavoina (1975; see DeMaris, 2002), the

Table 6. Results per group of (1) simultaneous regression analyses on mean accuracy with word frequency, bigram frequency residuals (Bigram residuals), orthographic neighborhood size residuals (N_Size residuals), and word frequency of the most frequent orthographic neighbor (N_Freq residuals) as predictors and (b) binary logistic regression analyses on item instability scores (IS_{item}) with the same four predictors

	(a) Regression analyses: Accuracy		(b) Binary logistic regression analyses: IS_{item}							
Reading group			**Model statistics**					**Predictor statistics**		
Predictors	**R^2 adjusted**[a]	**Beta**	**Classification**	***p***	**−2LL**	**χ^2**	**R^{2a}**	**B (coef.)**	**Wald (1)**	**Odds Ratio**[b] **(Exp(B))**
TR1	$.03^{***}$		79.9%	.001	360.2	19.2	.08			
Word frequency		$.12^{**}$						0.6	9.4	1.7^{**}
Bigram residuals		−.02						0.3	1.6	1.4
N_Size residuals		$.10^{*}$						0.0	0.6	1.0
N_Freq residuals		.09						0.2	1.1	1.2
TR2	$.02^{**}$	85.8%	.98	132.0	0.4	.01				
Word frequency		$.14^{***}$						0.1	0.0	1.1
Bigram residuals		.03						−0.2	0.3	0.8
N_Size residuals		$.08^{*}$						0.0	0.3	1.0
N_Freq residuals		.01						0.1	0.2	1.2
PR1	$.05^{***}$		76.7%	.07	443.8	8.5	.03			
Word frequency		$.15^{***}$						0.2	2.8	1.3
Bigram residuals		−.02						−0.0	0.0	1.0
N_Size residuals		$.19^{***}$						0.6	6.0	1.1^{*}
N_Freq residuals		.01						−0.2	0.7	0.9

Table 6. (*continued*)

	(a) Regression analyses: Accuracy		(b) Binary logistic regression analyses: IS_{item}							
PR2	.03***		83.9%	.03	209.1	10.8	.07			
Word frequency		.16***						0.6	6.9	1.9**
Bigram residuals		−.06						−0.1	0.1	0.9
N_Size residuals		.02						−0.0	0.8	1.0
N_Freq residuals		.13**						0.4	1.4	1.4

Note. TR1 = Typical Readers in Grade 1, TR2 = Typical Readers in Grade 2, PR1 = Poor readers matched to the reading level of TR1, PR2 = Poor Readers matched to the reading level of TR2.

a McKelvey and Zavoina's (1975) pseudo R^2 measure for logistic regression

b (p / 1 - p).

* $p < .05$

** $p < .01$

*** $p < .001$.

estimated amount of variance explained by the regression of IS_{item} on word characteristics was found to be low, from .03 for PR1 to .08 for TR1. The strongest predictor of accuracy, Word frequency, was also the strongest predictor of IS_{item} for TR1 and PR2: the higher a word's frequency, the greater the instability of the reading errors. IS_{item} was predicted by Neighborhood size residuals for PR1. In summary, for all but the most competent readers in our study (TR2), the instability of reading errors could be explained to a small degree by word characteristics and most strongly by Word frequency: The higher a word's frequency, the less stable the reading errors.

4. Discussion

The instability of reading errors in grade 1 and 2 typical readers (TR1 and TR2, respectively) and reading-level matched poor readers (PR1 and PR2, respectively) was explored using Dutch CVC words that are orthographically fully transparent in reading and thus can be read by applying grapheme-phoneme correspondence rules. Selected words were a representative sample of word tokens.

Overall reading accuracy was high and stable for all groups of children. This finding is in keeping with the CVC accuracy scores reported for a speeded reading task by Verhoeven and van Leeuwe (2009). In our Dutch study a larger percentage of errors was unstable than in the English study by Gough et al. (1992). Possibly, errors in a transparent orthography as Dutch are characterized by a higher instability than errors in an opaque orthography as English. Ellis et al. (2004) show that differences in orthographic transparency result in a different nature of reading errors (see also Guron & Lundberg, 2004). Readers of opaque orthographies tend to recognize words on the basis of partial visual analysis, whereas readers of transparent orthographies synthesize pronunciations by means of decoding. The different reading strategies in opaque versus transparent orthographies may also entail differences in error instability.

The Instability Score, which corrects for the number of errors, shows the children's reading errors to be not completely stable or unstable. The reading errors of the second graders were more unstable than the errors of the first graders. Thus, our data show that error *stability* is associated with an early phase of a decoding strategy under acquisition. This is in line both with the basic principle of miscue analyses, that recurrent, similar errors (error patterns), provide a window into strategies in progress, and with previous research on error instability (Steenbeek-Planting et al., 2013a).

The poor readers' errors did not differ in their instability from the typical readers' errors. The reading behavior of the poor readers evidently does not stand out by a larger random component. Note, however, that our comparison of poor and typical readers involved groups matched with respect to reading level and thus of different ages. Do poor readers differ from age-matched typical readers? In an additional analysis, the error instability of the 40 children in the PR1 group was compared to that of 40 age-matched typical readers from grade 3 (mean age: 9;3, *SD*: 0;4). The instability of the errors produced by the poor readers was significantly *lower* than that of the age-matched typical readers ($F(1, 65) = 12.84$, $p = .001$, $\eta_p^2 = .17$). That this difference was not found in the comparison with typical readers *matched for reading level*, suggests that the reading development of Dutch poor readers with respect to error instability is delayed as opposed to deviant.

The instability of reading errors was predicted to some extent by the words' lexical and sublexical characteristics. Word frequency was an important predictor of our index of error instability: the higher a word's frequency, the higher the degree of (corrected) error instability. Probably, the high frequency words can be identified by both visual access and by using grapheme-phoneme correspondence rules. As we assume discussing the Dual-Route Cascaded model further on, errors in these words that are presumably lexicalized, are due to inattentiveness or other stochastic processes. While low frequency words need to be identified by grapheme-phoneme correspondence rules and to a lesser degree via direct lexical access and involve errors that are rule based. The effect for word frequency was found for all but the most competent readers. Presumably, their self-teaching (Share, 1995) has matured to such a level that they have reached ceiling in reading both novel and high-frequency words, and the instability of their reading errors does not longer depend upon a word's frequency.

In the sublexical word characteristics studied, neighborhood size predicted error instability for PR1 only: the larger the neighborhood size, the higher the instability score. Thus, errors made in words with few neighbors were characterized as rather stable. This suggests that the instability of early reading errors is associated with the use of sublexical word units. We do not find this association in TR1, which suggests qualitative differences between beginning poor and typical readers. It may also be the case that poor readers matched to typical grade 1 readers tend to use an orthographic analogy strategy for the reading of both novel and known words (Wood, 2002).

In terms of the Dual-Route Cascaded model (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001) the orthographically transparent CVC words can be identified by both visual access and by using grapheme-phoneme correspondence rules. We hypothesize that the readers at level 1 have *not yet* fully acquired the relevant grapheme-phoneme correspondence rules. Chances are that they apply improper rules to words that are not lexicalized, with reading errors as a result. Lack of grapheme-phoneme correspondence knowledge leads to errors that are 'rule-governed' and thus rather stable reading errors. With increasing reading experience, an increasing number of items will be identified via direct retrieval from the visual mental lexicon. The reading errors of the more experienced readers at level 2, therefore, probably, stem from inattentiveness and other stochastic processes in retrieving information from the visual mental lexicon (confusing with look-alikes, for example).

Our results can also be interpreted partly in more plain and general terms. The beginning, grade 1, readers decode words consciously, they often are not aware of making errors, and they might even be convinced that their reading is correct (illusion of knowing, Serra & Metcalfe, 2009), resulting in relatively many stable errors compared to the grade 2 readers. As reading competence increases, automatic and attentionless reading increases. As a result, random behavior is a relatively greater cause for making errors in grade 2 than in grade 1, resulting in a relatively high number of unstable errors.

The word selection was based on token-count sampling and reflects the reading of words children generally encounter in print, but it does not represent the full range of words that comprise the Dutch vocabulary. Moreover, as error instability correlates with word frequency and our results are based on a sample of predominantly high-frequency words, our results might be biased (see also Share, 1995). To determine whether the degree of error instability is affected by the characteristics of the word set, an additional experiment was conducted with CVC words sampled from a *type* list (henceforth type-list experiment). Random sampling – according to Zipf's law (Zipf, 1936) – results in a sample with a relatively small number of high-frequency words and a high number of low-frequency words (see also Baayen, 2001). The same methodology and procedures were followed as in the main, token-count experiment.

On the whole, the results of the token-count experiment were replicated in the type-list experiment:[7] The instability of the children's reading errors (IS_{subj}) was similar and the same between-group differences were observed. As in the token-count experiment, accuracy was only partially predicted by the word characteristics, and word frequency was again found to be the strongest predictor. However, the amount of variance in accuracy and error instability explained by the different word factors was rather low in the type-list experiment. Using the terminology of the Dual-Route Cascaded model, we suggest that the words in the token-count experiment had a higher chance of being read via the lexical route. Reading via the lexical route entails that words are recognized as a visual whole, implying that

---

[7] Details of the type-list experiment are published in the doctoral dissertation of the first author.

word characteristics such as orthographic neighborhood features and bigram frequency come into play more than in reading via the nonlexical route.

A limitation of our study is that it is focused on regular CVC words, and thus on a limited section (9.60%) of the Dutch lexicon. Further research should clarify whether our results apply as well to words that involve the application of contextual, graphotactical or morphological rules, or to words with an idiosyncratic spelling.

Our results have several implications. In contrast to test scores, item scores in both poor and typical readers vary between test occasions. A theoretical implication is that models to explain the ability of a reader to successfully read *a specific word,* should also account for the low reliability of this skill.

Our results also have theoretical implications for the field of research on metacognitive monitoring that studies the relationship between task performance and a judgement about that performance (see e.g., Boekaerts & Rozendaal, 2010; Efklides, 2008; Winne & Nesbit, 2009). Typically a test answer is studied (correct or incorrect) and one's judgement about the answer (being correct or incorrect), resulting in a 2 x 2 matrix. We showed that the instability of test answers is positively related to level of fluency in a certain domain. Also, it has been shown that domain knowledge and skills are important determinants of metacognitive monitoring (Gutierrez, Schraw, Kuch, & Richmond, 2016; Tricot & Sweller, 2014), therefore, future research should clarify the relationship between error instability and domain-specific and domain-general metacognivie monitoring skills and its developmental trajectory.

A practical implication for assessment and intervention is that individual word reading errors should be interpreted with considerable caution as performance at the item-level varies over time. Reading errors are far from consistent and error instability is related to the level of reading competence of the student. This implicates that both the student's reading level and the instability of his or her reading errors should be taken into consideration for exercises in reading. Our study casts doubts on indiscriminately concentrating on specific errors in instruction and practice (see Steenbeek-Planting, van Bon, & Schreuder, 2012, 2013b).

Acknowledgements

We are grateful to Jan van Leeuwe, Giovanni ten Brink, Theo van der Weege, and Pieter van Groenestijn for their assistance in data computations. We also thank Jorita Bruining, Marjon van der Looij, and Linda van den Broek for their assistance in data collection.

References

Allison, P. D. (1999). *Logistic regression using the SAS System: Theory and Application*. Cary, NC: SAS Institute Inc.

Arduino, L. S., & Burani, C. (2004). Neighborhood effects on nonwords visual processing in a language with shallow orthography. *Journal of Psycholinguistic Research* 33: 75–95. doi: 10.1023/B:JOPR.0000010515.58435.68

Aro, M., & Wimmer, H. (2003). Learning to read: English in comparison to six more regular orthographies. *Applied Psycholinguistics* 24: 621–635. doi: 10.1017/S0142716403000316

Baayen, R. H. (2001). *Word frequency distributions*. Dordrecht, The Netherlands: Kluwer Academic Publishers. doi: 10.1007/978-94-010-0844-0

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1993). *The CELEX lexical database* [Computer Software]. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Balota, D. A., Yap, M. J., & Cortese, M. J. (2006). Visual word recognition: The journey from features to meaning (a travel update). In M. J. Traxler & M. A. Gernsbacher (eds.), *Handbook of psycholinguistics* (2nd ed.), 285–375. London: Elsevier. doi: 10.1016/B978-012369374-7/50010-9

Belsley, D. A. (1991). *Conditioning diagnostics: Collinearity and weak data in regression*. New York, NY: Wiley.

Boekaerts, M., & Rozendaal, J. S. (2010). Using multiple calibration measures in order to capture the complex picture of what affects students' accuracy of feeling of confidence. *Learning and Instruction* 20: 372–382. doi: 10.1016/j.learninstruc.2009.03.002

Booij, G. (1995). *The phonology of Dutch*. Oxford: Clarendon Press.

Chatterjee, S., Hadi, A. S., & Price, B. (2000). *Regression analysis by example*. New York, NY: Wiley.

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: 204–256. doi: 10.1037/0033-295X.108.1.204

Cossu, G., Shankweiler, D., Liberman, I. Y., & Gugliotta, M. (1995). Visual and phonological determinants of misreadings in a transparent orthography. *Reading and Writing: An Interdisciplinary Journal* 7: 237–256. doi: 10.1007/BF02539523

de Jong, P. F., & van der Leij, A. (2003). Developmental changes in the manifestation of a phonological deficit in dyslexic children learning to read a regular orthography. *Journal of Educational Psychology* 95: 22–40. doi: 10.1037/0022-0663.95.1.22

DeMaris, A. (2002). Explained variance in logistic regression: A Monte Carlo study of proposed measures. *Sociological Methods & Research* 31: 27–74. doi: 10.1177/0049124102031001002

Efklides, A. (2008). Metacognition: defining its facets and levels of functioning in relation to self-regulation and co-regulation. *European Psychologist* 13: 277–287. doi: 10.1027/1016-9040.13.4.277

Ellis, N. C., Natsume, M., Stavropoulou, K., Hoxhallari, L., van Daal, V. H. P., Polyzoe, N., Tsipa, M. L., & Petalas, M. (2004). The effects of orthographic depth on learning to read alphabetic, syllabic, and logographic scripts. *Reading Research Quarterly* 39: 438–468. doi: 10.1598/RRQ.39.4.5

Eurybase (2008). *Organization of the education system in the Netherlands*. Retrieved from http://eacea.ec.europa.eu/education/eurydice/eurybase_en.php

Frauenfelder, U. H., Baayen, R. H., Hellwig, F. M., & Schreuder, R. (1993). Neighborhood density and frequency across languages and modalities. *Journal of Memory and Language* 32: 781–804. doi: 10.1006/jmla.1993.1039

Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General* 113: 256–281. doi: 10.1037/0096-3445.113.2.256

Goikoetxea, E. (2006). Reading errors in first- and second-grade readers of a shallow orthography: Evidence from Spanish. *British Journal of Educational Psychology* 76: 333–350. doi: 10.1348/000709905X52490

Goswami, U. (2002). Phonology, reading development, and dyslexia: A cross-linguistic perspective. *Annals of Dyslexia* 52: 1–23. doi: 10.1007/s11881-002-0010-0

Gough, P. B., Juel, C., & Griffith, P. L. (1992). Reading, spelling, and the orthographic cipher. In P. B. Gough, L. C. Ehri, & R. Treiman (eds.), *Reading acquisition*, 35–48. Hillsdale, NJ: Lawrence Erlbaum Associates.

Guron, L. M., & Lundberg, I. (2004). Error patterns in word reading among primary school children: A cross-orthographic study. *Dyslexia* 10: 44–60. doi: 10.1002/dys.260

Gutierrez, A. P., Schraw, G., Kuch, F., & Richmond, A. S. (2016). A two-process model of metacognitive monitoring: Evidence for general accuracy and error factors. *Learning and Instruction* 44: 1–10.

Habib, M. (2000). The neurological basis of developmental dyslexia: An overview and working hypothesis. *Brain* 123: 2373–2399. doi: 10.1093/brain/123.12.2373

Landauer, T. K., & Streeter, L. A. (1973). Structural differences between common and rare words: Failure or equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Verbal Behavior* 12: 119–131. doi: 10.1016/S0022-5371(73)80001-5

Landerl, K. (2001). Word recognition deficits in German: More evidence from a representative sample. *Dyslexia* 7: 183–196. doi: 10.1002/dys.199

McKelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal dependent variables. *Journal of Mathematical Sociology* 4: 102–120. doi: 10.1080/0022250X.1975.9989847

McKenna, M. C., & Picard, M. C. (2006). Revisiting the role of miscue analysis in effective teaching. *The Reading Teacher* 60: 378–380. doi: 10.1598/RT.60.4.8

Murray, W. S., & Forster, K. I. (2004). Serial mechanisms in lexical access: The rank hypothesis. *Psychological Review* 111: 721–756. doi: 10.1037/0033-295X.111.3.721

Nunnally, J. C. & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed). New York, NY: McGraw-Hill.

Ognjenović, V., Lukatela, G., Feldman, L. B., & Turvey, M. T. (1983). Misreadings by beginning readers of Serbo-Croatian. *Quarterly Journal of Experimental Psychology* 35A: 97–109. doi: 10.1080/14640748308402119

Patel, T. K., Snowling, M. J., & de Jong, P. F. (2004). A cross-linguistic comparison of children learning to read in English and Dutch. *Journal of Educational Psychology* 96: 785–797. doi: 10.1037/0022-0663.96.4.785

Rastle, K. (2007). Visual word recognition. In: M. G. Gaskell (ed.), *The Oxford Handbook of Psycholinguistics*, 71–87. New York, NY: Oxford University Press.

Serra, M. J., & Metcalfe, J. (2009). Effective implementation of metacognition. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (eds.), *Handbook of metacognition in education*, 278–298. Mahwah, NJ: Erlbaum.

Serrano, F., & Defior, S. (2008). Dyslexia speed problems in a transparent orthography. *Annals of Dyslexia* 58: 81–95. doi: 10.1007/s11881-008-0013-6

Seymour, P. H. K., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology* 94: 143–174. doi: 10.1348/000712603321661859

Share, D. L. (1995). Phonological recoding and self-teaching: *Sine qua non* of reading acquisition. *Cognition* 55: 151–218. doi: 10.1016/0010-0277(94)00645-2

Steenbeek-Planting, E. G., van Bon, W. H. J., & Schreuder, R. (2012). Improving word reading speed: Individual differences interact with a training focus on successes or failures. *Reading and Writing: An Interdisciplinary Journal* 25: 2061–2089. doi: 10.1007/s11145-011-9342-7

Steenbeek-Planting, E. G., van Bon, W. H. J., & Schreuder, R. (2013a). Instability of children's reading errors in bisyllabic words: The role of context-sensitive spelling rules. *Learning and Instruction* 26: 59–70. doi: 10.1016/j.learninstruc.2013.01.004

Steenbeek-Planting, E. G., van Bon, W. H. J., & Schreuder, R. (2013b). Improving the accuracy of reading bisyllabic words that involve context-sensitive spelling rules: Focus on successes or on failures? *Reading and Writing: An Interdisciplinary Journal* 26: 1437–1458. doi: 10.1007/s11145-012-9425-0

van Bon, W. H. J. (2007). *De Doorstreepleestoets* [Paper-and-pen lexical decision task]. Leiden, The Netherlands: PITS.

van Bon, W. H. J., Bouwmans, M., & Broeders, I. N. L. D. C. (2006). The prevalence of poor reading in Dutch special elementary education. *Journal of Learning Disabilities* 39: 482–495. doi: 10.1177/00222194060390060101

van Bon, W. H. J., Hoevenaars, L. T. M., & Jongeneelen, J. J. (2004). Using pencil-and-paper lexical-decision tests to assess word decoding skill: Aspects of validity and reliability. *Journal of Research in Reading* 27: 58–68. doi: 10.1111/j.1467-9817.2004.00214.x

van den Bos, K. P., Lutje Spelberg, H. C., Scheepstra, A. J. M., & de Vries, J. R. (1994). *De Klepel – Verantwoording, Diagnostiek en Behandeling* [Nonword Reading Test]. Nijmegen, The Netherlands: Berkhout.

Verhoeven, L. T. W. (1995). *Drie-Minuten-Toets* [Three-Minutes Test]. Arnhem, The Netherlands: Cito.

Verhoeven, L. T. W., & van Leeuwe, J. (2003). Ontwikkeling van decodeervaardigheid in het basisonderwijs [Development of decoding ability in primary education]. *Pedagogische Studiën* 80: 257–271.

Verhoeven, L. T. W., & van Leeuwe, J. (2009). Modeling the growth of word-decoding skills: Evidence from Dutch. *Scientific Studies of Reading* 13: 205–223. doi: 10.1080/10888430902851356

Wimmer, H., & Goswami, U. (1994). The influence of orthographic consistency on reading development: Word recognition in English and German children. *Cognition* 51: 91–103. doi: 10.1016/0010-0277(94)90010-8

Winne, P. H., & Nesbit, J. C. (2009). Supporting self-regulated learning with cognitive tools. In D. J. Hacker, J. Dunlosky, & A. C. Grasser (eds.), *Handbook of metacognition in education*, 258–277. Mahwah, NJ: Erlbaum.

Wood, C. (2002). Orthographic analogies and phonological priming effects. *Journal of Research in Reading* 25: 144–159. doi: 10.1111/1467-9817.00165

Zipf, G. K. (1936). *The psycho-biology of language: An introduction of dynamic philology*. London: Routledge.

Appendix A. The instability score (IS)

$$IS = \frac{IS_{obs} - IS_{min}}{IS_{max} - IS_{min}}$$

IS_{obs} denotes the observed number of unstable errors between Time 1 and Time 2. IS_{min} designates the minimum number of unstable errors ($|e_1 - e_2|$); e_1 is the number of errors at Time 1, e_2 is the number of errors at Time 2. IS_{max} denotes the maximum number of unstable errors and is calculated as ($n - | e_1 + e_2 - n|$); n is the number of words that are read on both occasions. When a child thus reads 100 words on two occasions and misidentifies 30 words at Time 1 and 20 words at Time 2, for example, the maximum number of unstable errors is 50 when all of the misidentifications involve different words and no word is thus read erroneously on both occasions. Because only 20 words can be misidentified on both occasions, 10 words *must* have been misidentified on one occasion but not the other (i.e., unstably). Thus the minimum number of unstable errors, that is words misidentified at Time 1 but not at Time 2, is 10.

Address for correspondence

Esther G. Steenbeek-Planting
Behavioral Science Institute
Radboud University Nijmegen
Department of Special Education
P.O. Box 9104
6500 HE Nijmegen
The Netherlands

e.steenbeek@pwo.ru.nl

Co-author details

Wim H. J. van Bon
Behavioral Science Institute
Radboud University Nijmegen
P.O. Box 9104
6500 HE Nijmegen
The Netherlands