edual editorial UNIVERSIDAD DE ALMERÍA

Electronic Journal of Research in Educational Psychology
ejrep

UNIVERSIDAD DE ALMERÍA

Ilustre Colegio Oficial de Psicología DE ANDALUCÍA ORIENTAL

Psychometric properties of a silent word reading test (LEO-1-min)

Edurne Goikoetxea[1], Wim Van Bon[2], Gorka Fraga[3] y Naroa Martínez[1]

[1] Department of Psychology and Education, University of Deusto, Bilbao
[3] Faculty of Social Sciences, Radboud University, Nijmegen
[3] Department of Developmental Psychology, University of Zurich, Zurich

España, Los Países Bajos, Suiza

Correspondencia: Edurne Goikoetxea. Av. de las Universidades, 24. 48007 – Bilbao, Vizcaya, SPAIN. E-mail: egoikoetxea@deusto.es

Abstract

Introduction. Teachers and researchers often need to evaluate word decoding skill in group-wise and in a short time. The LEO-1-min test is created to measure word reading through a lexical decision procedure where the examinee identifies pseudowords in a list of frequent words.

Objetive. To examine the reliability and validity of LEO-1-min, a silent word reading test, suitable for quick assess of reading abilities in a wide age range of students.

Method. Participants were 284 children from 1st to 6th grade of a subsidized Primary School. We created four alternate forms of the LEO-1-min, each with 180 stimuli (132 words and 48 pseudowords).

Results. The results show an adequate parallel forms reliability of the scores (range *r*s = from .57 to .81). High correlations were found between the scores on the LEO-1-min and the scores on a standardized reading aloud test. The discriminant analysis of the scores on the LEO–1-min shows a high level of success in predicting the oral word decoding performance.

Discussion and Conclusion. LEO-1-min reliability is acceptable to good. Lexical decision in LEO-1-min and oral reading are highly correlated, which support using lexical decision as a groupwise test to screen for poor word readers. Form A of the test and provisional scales are presented for each primary grade.

Keywords: Word decoding, reading development, reading test, pencil-and-paper test, lexical decision task.

Resumen

Introducción. Maestros e investigadores necesitan con frecuencia evaluar la lectura de palabras en grupo y en poco tiempo. El test LEO-1-min fue creado para medir la lectura de palabras a través de una tarea de decisión léxica donde el examinado debe identificar pseudopalabras dentro de una lista de palabras frecuentes.

Método. Participaron 284 niños de 1º a 6º de una escuela pública concertada. Se crearon cuatro formas alternativas del LEO-1-min, cada una con 180 estímulos (132 palabras y 48 pseudopalabras).

Resultados. Los resultados mostraron que una adecuada fiabilidad para formas alternas (rango rs = de .57 a .81). Se encontraron altas correlaciones entre las puntaciones en el LEO-1-min y las de un test estandardizado de lectura en voz alta. El análisis discriminante de las puntuaciones del LEO-1-min mostró un alto nivel de éxito en la predicción del rendimiento en lectura oral.

Discusión y **Conclusion**: La fiabilidad de las puntuaciones del LEO-1-min es aceptable a buena. La tarea de decisión léxica del LEO-1-min mostró una alta correlación con la lectura oral, lo que apoya el uso de la decisión léxica como test grupal para identificar rápidamente lectores con pobre lectura de palabras. Se ofrece la Forma A del test y baremos provisionales por curso.

Palabras clave: Lectura de palabras, desarrollo lector, test de lectura, tareas de decisión léxica.

Introduction

Reading teachers use a wide variety of teaching and assessment strategies (Lacina & Block, 2011). The use of brief and individual reading aloud tests is a frequent part of the class routine in order to supervise reading progress throughout the school year. If such tests are adequately organised within the teaching routine, they offer a maximum benefit at a minimum cost for the child and the school. The reason is that a teacher who is knowledgeable is usually more efficient in detecting problems than many tests are (Wilson & Jungner, 1968). In addition, teachers can establish objectives and adapt their teaching without delay if they perceive that the child's progress is insufficient (Förster & Souvignier, 2015; Förster, Kawohl, & Souvignier, 2018). However, routinely applying reading aloud tests is difficult when the number of children is very large, or when repeated assessment is needed to follow the progress of students and quickly identify those who are at risk of failing.

Before submitting your newly formatted article, please reread it in its entirety from the perspective of someone not from your country. Elements that pertain to your country's educational system may need to be explained for the larger audience.

In order to perform quick and frequent reading assessments of large groups of children, it is necessary to have scientifically sound tests, ideally with alternative forms. In Spain, there are well-founded tests to evaluate reading (including word decoding) in Primary school (Cuetos, Rodríguez, Ruano, & Arribas, 2007; Defior et al., 2006; Jiménez, Gove, Crouch, & Rodríguez, 2014) and brief tests that measure specific aspects of reading, such as reading fluency (González-Trujillo, Calet, Defior, & Gutiérrez-Palma, 2014) but their individual administration makes them time-consuming. The only brief test for groups we know evaluates reading comprehension (Marín & Carrillo, 1999).

We do not know of any brief group tests to measure word identification or word reading, that is, the retrieval of a word's phonology and meaning, even though these skills strongly determine reading comprehension (Jenkins, Fuchs, Van den Broek, Espin, & Deno, 2003; Kim, Petscher, Schatschneider, & Fooman, 2010; Klauda & Guthrie, 2008; Perfetti & Stafura, 2014). Poor word reading triggers a vicious circle that limits the progression to a level of reading for meaning and learning. Children who struggle with word identification have difficulty reaching reading fluency. Their reading stays laborious, and their motivation to read,

their reading practice, and their reading comprehension stays low (Rasinski, Reutzel, Chard, & Linan-Thompson, 2011). In fact, research on what comprehension tests really measure shows that the variability in this measure is mainly due to word reading skill (Keenan & Meenan, 2014).

The assessment of word reading skill, whether by paper and pencil or computerized, is mainly performed with two tasks: naming (oral reading or reading out loud) or lexical decision (deciding whether a letter string is an existing word or not). The evidence about the validity of these tasks comes from studies that have compared them to silent reading, specifically with eye fixation times. Eye fixation times could be considered the gold standard because they are registered during natural skilled reading, and they have been shown to reflect cognitive reading processes (e.g., Foster, Ardoin, & Binder, 2018; Rayner & Reichle, 2010; Rayner, Sereno, Morris, Schmauder, & Clifton, 1989). Studies with English-speaking subjects reveal that, although the task of reading out loud is more similar to eye fixation times, the lexical decision task also has a strong correlation with eye fixation times, which demonstrates that both tasks, naming and lexical decision, offer valid data to estimate silent reading (Schilling, Rayner, & Chumbley, 1998; also see Forster, 1976; Forster & Chambers, 1973). Another question is whether the previous results are replicated in different orthographies. Data of eye movements are lacking for Spanish, a shallow orthography with simple syllabic structure compared with English, but Kuperman, Drieghe, Keuleers, and Brysbaert (2013) show that Dutch (modestly shallow) and English (very deep) data presents the same pattern of correlations between eye movement, lexical decision, and naming latencies. Interestling, these authors find a different pattern than the precedent research: there is a stronger correlation between of eye movement latencies with lexical decision than with naming lantecies.

The test introduced in this study is based on previous work. First our experience with lexical decision tasks on computers (Goikoetxea & Ferrero, 2019). Although the computer allows recording the response time and seems more scalable than paper and pencil assessment, in fact it imposes a limitation on the number of children simultaneously evaluated due to the large number of errors made by children while working with computers (Moret-Tatay & Perea, 2011). Therefore, we decided to create a paper-and-pencil test. Second, the LEO-1-min is based on a test that measures word reading in the Dutch language: the Doorstreepleestoets or paper-and-pencil lexical decision task by Van Bon (2007; see also Van Bon, Hoevenaars, & Jongeneelen, 2004; van Bon & Libert, 1997; Van Bon, Tooren, & Van

Eekelen, 2000). This test was developed to screen poor readers in 2nd and 3rd grade of primary school in need of support measures or reinforcement. The lexical-decision task, performed in 1 minute, demonstrated a good correlation between the parallel forms (mean of the rs: 0.82 for 2nd grade and 0.66 for 3rd grade), indicating good score reliability. A low correlation with a symbols test (range of the rs: from 0.12 to 0.25 for 2nd grade and from 0.07 to 0.20 for 3rd grade) indicates that motor skills were not decisive in the performance. It also showed criterion validity due to the high correlations with the scores on a standardized test of reading aloud (mean of the rs: 0.79 for 2nd grade and 0.67 for 3rd grade). These results show that this test was appropriate to measure the reading of 2nd and 3rd grade children. Therefore, we have based the construction of our new test on the Doorstreepleestoets.

Objectives and Hypotheses

The goals of the present study were, first, to create a word reading test using this lexical decision procedure, called the LEO-1-min, with four alternative forms. According to the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (2014), two or more versions of a test that are considered interchangeable are referred to as alternate (or equivalent) forms. Accordingly, we designed the four forms of the LEO-1-min to have the same general distribution of content, item formats, and administration procedure. Although the alternate forms do not completely eliminate the effects of practice or memory, they do reduce these effects because the items are not identical. Based on these considerations, we expect the four forms of the LEO-1-min to have approximately the same score means and standard deviations in each grade. The second goal was to investigate reliability and convergent-discriminant validity of LEO-1-min. Based on the process of construction, we expect that parallel reliability would be satisfactory, the LEO-1-min would correlate higher with an oral reading test than with an oral arithmetic test, and that a discriminant analysis would reveal that scores on LEO-1-min predicts the judgments of teachers about the reading level of their students. The LEO-1-min is designed to assess reading in primary education, lower secondary, and high school, but in this paper, we only show the psychometric properties for forms A and B in a sample from 1st to 6th grades of primary education, and for forms C and D in a sample of three primary education grades.

Method

Participants

In all, 284 children in 1st to 6th grades from a subsidized primary school in Guecho, Vizcaya took part in the experiment: 48 in Grade 1 (31.3% girls; 35.4% missing data; mean age = 5.94, *SD* = 0.24), 47 in Grade 2 (51% girls, 21% missing data; mean age = 7,0; *SD* = 0,21), 51 in Grade 3 (14% girls; 72% missing data; mean age = 7.96; *Dt* = 0.20), 42 in Grade 4 (31% girls, 33% missing data; mean age = 8.98; *Dt* = 0.27), 50 in Grade 5 (18% girls; 70% missing data; mean age = 10.08; *Dt* = 0.40), 46 in Grade 6 (33% girls, 33% missing data; mean age = 11.00; *Dt* = 0.30). The school serves a middle class population. Additionally, seventeen participants were excluded from the analyses due to errors in the use of the test. One child showed the need for special education. His/her data were also excluded from the calculations in the following. The sampling was incidental. The mother tongue and first language in school was Spanish, making up 65% of the class hours, with the remaining 35% in Basque and English. An alphabetic method was used to teach reading and writing, where the names of the letters and their sounds were learned in combination with each of the 5 vowels.

Instruments

LEO-1-min is a speed test designed to determine visual word recognition fluency in readers in a wide range of ages and reading competence. Children are asked to identify the pseudowords among the real words in a list of items, for one minute.

The four alternate forms of the LEO-1-min, each with 180 stimuli, were randomly drawn from a pool of 528 most frequent nouns (excluding names and colloquialisms; 32 words of one syllable, 279 of two syllables, 166 of three syllables, 43 of four syllables, and 8 of five syllables) selected from the vocabulary of LEXIN (Corral, Ferrero, & Goikoetxea, 2009). The list of 528 words was used to create the pseudowords by substituting interior letters according to the length of the words. In mono-syllabic words, the last letter of each word were changed (e.g. gas [gas] to gar), In dysillabic words, the first consonant in the second syllable was substituted (e.g., mesa [table] to meda). In words with 3 or more syllables, the first consonant in the second and third syllables was substituted (e.g., mañana (tomorrow) to mavaga). The syllabic structure of the words was preserved.

Next, we examined whether the alternative forms were similar in psycholinguistic variables with known effects on the visual word recognition in Spanish: lexical frequency

(Defior, Justicia, & Martos, 1996), also controlled in previous studies (Van Bon, Hoevenaars, & Jongeneelen, 2004), length (Acha & Perea, 2008), neighborhood (Perea & Rosa, 2000), and syllabic frequency (Carreiras, Álvarez, & de Vega, 1993). Each list included 132 nouns. The mean frequency of these nouns is 66.77 (range: 37.13-101.4) in LEXIN and 119.56 (range: 30.89-850.89) in LEXESP (Sebastián-Gallés, Martí, Carreiras, & Cuetos, 2000). The mean length is 5.19 letters (range: 3-12), and the mean neighbourhood is 3.39 (range: 0-23). The mean syllabic frequency of the nouns is 876.97 (range: 73.45- 6737.32) in the first syllable, 1289.43 (range: 34.46- 9665.36) in the second syllable, 1682.04 (range: 48.57- 9600.36) in the third syllable, and 2008.47 (range: 51.43- 6210.18) in the fourth syllable in Buscapalabras (Davis & Perea, 2005). Each list included a proportion of three words to each pseudoword, as in the previous test for primary (Van Bon et al., 2000, 2004), yielding lists of 180 stimuli: 132 nouns and 48 pseudowords in random order.

In each list, the stimuli are arranged in four columns. The font is Arial 13, single-spaced. The pseudowords are distributed throughout the four columns, and at least one pseudoword appears in the lower half of the last column, and never as the first stimulus in the first column. The students are required working column by column, to underline the pseudowords. After having done this for one minute, they put an X next to the last word seen. The instructions and stimuli on form A of the LEO-1-min are in Appendix A. The complete test with the four forms can be obtained for free by requesting a copy directly from the first author or from PsycTESTS (Database of the American Psychological Association). The Symbols test was created to estimate the influence of perceptual-motor components of the underlining task on the scores on the LEO-1-min. The list included 180 strings of letters made up of the same letter of the alphabet (except the X), repeated a minimum of 3 times and a maximum of 12. In 48 of the 180 strings, one or more letters were replaced by X. The task required underlining the strings containing one or more Xs, until the experimenter said 'Stop'. Score was the number of items correctly completed within one minute.

Reading words and pseudowords subtests from the revised Battery for the Evaluation of Reading Processes in Primary (PROLEC-R; Cuetos et al., 2007), one of the tests most frequently used in Spain to measure oral word reading. The word and pseudoword reading subtests require the child to read aloud a list of words and pseudowords, respectively, and the accuracy and time are measured. The internal consistencyreliability for the word and pseudoword reading scores in this sample were .81 and .83, respectively.

Arithmetic test from the Wechsler Intelligence Scale for Children IV (WISC-IV; Wechsler, 2005). The Arithmetic test consists of arithmetic problems that the child must mentally solve in a limited time. This subtest has different start points according age. The internal consistency reliability of the Arithmetic scores in this sample was 0.86.

Procedure

The school administration and the adults responsible for the participants gave their informed consent to conduct the experiment. The children were tested at the school by two of the authors and three collaborators previously trained to perform the task. Each grade from 1st to 6th consisted of two classes. All the participants completed two of the four forms of the LEO-1-min and the symbols task. The form A and the symbols task were administered to all the participants, form B to half of the sample, that is, to the students in one classroom per grade, form C to the students in one classroom from 1st, 3rd and 5th grades, and form D to the students in one classroom from 2nd, 4th and 6th grades. The administration of both forms was performed in the same session with an interval of a few minutes. Due to practical limitations, the order of application of the two forms was not counterbalanced; form A was first when the parallel form was B, and the opposite was true when the parallel forms were C and D, always ending with the symbols task. The instructions for symbols were the same as for the LEO-1-min, but considering the target stimuli that included X.

To study the validity, we asked the classroom teachers to classify three children in each classroom in three categories: 1 for "very good reader", 2 for "normal reader", and 3 for "poor reader". The teachers thus classified 108 children from 1st to 6th grades (6 grades x 2 classrooms x 3 categories x 3 children). This subsample completed the PROLEC-R test and the Arithmetic test in a counterbalanced order and following a "blind" procedure (i.e., without knowing the data from the LEO-1-min or the classification made by the teachers).

Statistical analysis

For our main comparisons between test forms, grade and sex, we used a multivariate analysis of variance (MANOVA) on the different types of response, i.e., correct omissions, hits, false alarms and omissions. Tukey's post-hoc comparisons were used to further examine effects of interest. In order to test reliability, we used Pearson's correlations between the two forms of the LEO-1-min and, subsequently, we used repeated measures ANOVA with form as

a within–subject factor to examine difference between the grades. Pearson's correlations were also used in the convergent-discriminant analysis using LEO-1-min and PROLEC-R scores and the equivalence analysis using LEO-1-min scores and demographic variables. Finally, we performed predictive discriminant analysis using LEO-1-min scores as predictor and overall reading ability based on teachers judgments. Normality and equality of the variance-covariance matrix in this analysis were tested with Komogorov and Box's tests, respectively. Leave-One-Out method was used to estimate classification rate and Huberty's Z to statistically assess our classification.

All the statistical analyses were conducted with a Type I error probability set at .05. The analyses were performed using SPSS software, Version 21.0.

Results

Descriptive

The responses were labelled as "hits" for correctly underlining a pseudoword, "correct omissions" for real words that were not underlined, "false alarms" if real words were underlined, and "omissions" for pseudowords that were missed and not underlined (Figure 1). Next, the correct responses (total effectiveness) were calculated as the total number of correct omissions and hits (this is the same of, the number of attempted elements or the total number of items read by each child minus the omissions and the false alarms). The descriptive statistics for these types of responses on the LEO-1-min and the total scores on the symbols task are displayed in Table 1.

A MANOVA was conducted with types of response – total, correct omissions, hits, false alarms, and omissions – on each form of the LEO-1-min with grade and sex as independent factors. For Forms A and B, sex had no effect and did not interact with grade. Grade, however, had a statistically significant effect (see Table 1). Tukey's post-hoc comparisons revealed statistically significant differences at $p < .05$ between Grade 1 and all the other grades; Grade 2 and all the other grades; Grades 3 and 4 and all the other grades, but not between these two grades. This result can be explained for the normal decrease of children's reading growth after the first years of learning, but it could also be related to the low variability of Grade 4 scores in this sample. Grade 5 and 6 did not show any differences between them.

Table 1. *Total score on forms A and B of LEO-1-min and on the Symbols Task per grade.*

	Grade								
	1	2	3	4	5	6			
Form A	*n* = 46	*n* = 47	*n* = 51	*n* = 42	*n* = 50	*n* = 46	*F(5,283)*	*p*	η^2
Total score (0-180)	15.17 (10.75)	28.74 (11.59)	41.53 (13.09)	48.17 (11.71)	61.62 (17.55)	63.24 (15.71)	85.38	<.001	.61
Correct omissions (0-132)	9.43 (8.27)	19.26 (9.32)	29.53 (10.31)	34.40 (9.34)	45.82 (14.55)	46.35 (12.56)	81.40	<.001	.60
Hits (0-48)	5.74 (2.99)	9.49 (2.59)	12.00 (3.01)	13.76 (2.77)	15.80 (3.55)	16.89 (3.38)	83.23	<.001	.60
False alarms (0-132)	1.33 (2.03)	0.49 (0.72)	0.25 (0.56)	0.12 (0.33)	0.16 (0.37)	0.35 (0.60)	10.16	<.001	.16
Omissions (0-48)	1.41 (1.57)	1.36 (1.11)	1.55 (1.12)	1.48 (1.38)	2.26 (2.25)	1.35 (1.30)	2.65	.023	.05
	Grade								
Form B	*n* = 26	*n* = 24	*n* = 27	*n* = 20	*n* = 26	*n* = 23	*F*(5,145)	*p*	η^2
Total score (0-180)	14.54 (11.62)	25.92 (11.77)	41.15 (11.93)	43.20 (10.63)	53.12 (16.55)	57.04 (16.75)	36.12	<.001	.56
Correct omissions (0-132)	10.15 (8.15)	17.17 (8.32)	29.22 (9.24)	30.75 (8.30)	37.23 (11.61)	39.74 (11.20)	35.70	<.001	.56
Hits (0-48)	4.38 (3.79)	8.75 (8.33)	11.93 (3.10)	12.45 (2.70)	15.88 (5.04)	16.43 (4.91)	33.36	<.001	.54
False alarms (0-132)	1.38 (1.24)	0.88 (0.90)	0.52 (0.75)	0.05 (0.22)	0.31 (0.84)	0.22 (0.52)	8.64	<.001	.24
Omissions (0-48)	1.15 (1.64)	0.42 (0.58)	1.11 (1.53)	1.10 (1.17)	0.54 (0.95)	1.04 (1.22)	1.70	.14	.06
Form C	*n* = 20		*n* = 24		*n* = 23		*F*(5, 67)	*p*	η^2
Total score (0-180)	16.05 (8.64)		42.63 (15.04)		70.09 (21.53)		59.58	<.001	.65

Table 1 (continued)

	Grade								
	1	2	3	4	5	6			
Form A	n = 46	n = 47	n = 51	n = 42	n = 50	n = 46	F(5,283)	p	η2
Correct omissions (0-132)	10.55 (5.85)		29.83 (10.77)		49.83 (15.87)		59.77	<.001	.65
Hits (0-48)	5.50 (2.89)		12.79 (4.35)		20.39 (6.49)		50.00	<.001	.61
False alarms (0-132)	0.90 (0.97)		0.63 (0.65)		1.17 (1.23)		1.87	.163	.06
Omissions (0-48)	0.45 (0.69)		0.33 (0.48)		0.65 (0.93)		1.61	.163	.04
Form D		n = 23		n = 22		n = 24	F(2, 68)	p	η2
Total score (0-180)		32.26 (9.55)		50.95 (15.48)		65.88 (15.48)	36.45	<.001	.53
Correct omissions (0-132)		22.65 (7.25)		36.36 (10.01)		46.33 (11.01)	36.13	<.001	.52
Hits (0-48)		9.61 (2.41)		14.59 (4.91)		19.54 (4.63)	33.94	<.001	.51
False alarms (0-132)		0.78 (1.04)		0.55 (0.72)		1.08 (1.18)	1.65	.201	.05
Omissions (0-48)		1.26 (0.75)		1.00 (0.93)		0.67 (0.76)	3.15	.049	.09
Symbols	n = 46	n = 47	n = 51	n = 42	n = 50	n = 47	F(5,282)	p	η2
Total score (0-180)	48.33 (12.46)	63.34 (14.42)	78.04 (13.66)	83.07 (14.49)	102.06 (16.51)	108.40 (23.46)	92.29	<.001	.63

This pattern was identical for the total, correct omissions, and hits. However, Grade 1 was the only grade that made more false alarms and omissions than all the others probably because for a beginning reader a few letters suffice to recognize a word (false alarms) but have a limited lexicon (omissions). Similarly, for Forms C and D, grade had a significant effect (see Table 1), but sex did not. Tukey's post-hoc comparisons showed significant differences between Grades 1, 3, and 5 on Form C, and differences between Grades 2, 4, and 6 on Form D.

The pseudoword identification response (underlining) undeniably involves perceptual motor skills. But the more items children can handle in the same time on the symbols test, compared to the LEO-1-min, the more the LEO-1-min is driven by other factors apart from these perceptual-motor skills. We would, therefore, expect the significant difference that Table 1 suggests.

We calculated the correlations between the scores for words and symbols; a high correlation can indicate an important contribution of motor skills on both tasks. These correlations are shown in Table 1. The results show moderate and positive correlations between the performance on the LEO-1-min in most of the grades, except in 3rd and 4th grades, where the correlation decreases.

Parallel forms reliability

First, we calculated the Pearson's correlation between the scores on the two forms of the LEO-1-min. The results of this analysis, displayed in Table 2, show positive and moderate-to-high correlations throughout primary education.

To further examine the reliability of the scores on the alternate forms of the LEO-1-min, we compared the means on the forms of the test, looking for differences that would question their equivalence. A repeated-measures ANOVA with the form as the within-subject factor (two levels: A and B) and grade as between-subject factor compared the forms of the LEO-1-min. There were no statistically significant differences in performance between forms A and B in any grade, F (1,25)= 1.97, p = 0.173, η^2 = 0.073 in Grade 1, F (1,23)= 1.97, p = 0.173, η^2 = 0.079 in Grade 2, F (1,26)= 1.06, p = 0.313, η^2 = 0.039 in Grade 3, F (1,19)= 2.93, p = 0.103, η^2 = 0.134 in Grade 4, F (1,25)= 1 .14, p = 0 .296, η^2 = 0 .044 in Grade 5, F (1,22)= 1 .71, p = 0 .205, η^2 = 0 .072 in Grade 6. A second ANOVA with forms A and D showed no form differences either in any grade, F (1,22) = 2 .65, p = 0 .118, η^2 = 0 .107 in Grade 2, F (1,21) < 1 in Grade 4, and F (1,28) < 1 in Grade 6. A third ANOVA with forms A and C, however, showed significant form differences in two grades: Grade 1, F (1,19)= 4.84, p = 0.040, η^2 = 0.203 and Grade 3, F (1, 23)= 6.72, p = 0.016, η^2 = 0.226, but there were no differences in Grade 5, F (1,22) < 1.

Table 2

Pearson's Correlations between Forms A and other forms of LEO-1-Min and the Symbols Task

	Form A			Form B			Form C			Form D	
	All other forms	Symbols		Form A	Symbols		Form A	Symbols		Form A	Symbols
Grade	*n*	*r*	*r*	*n*	*r*	*r*	*n*	*r*	*r*	*n*	*r*	*r*
1	46	.81**	.40**	26	.84**	0.32	20	.85**	0.60**			
2	47	.76**	.51**	24	.84**	0.69**				23	.67**	.52*
3	51	.81**	.19	27	.65**	0.40*	24	.94**	0.20			
4	42	.57**	.27	20	.56**	0.26				22	.57**	.05
5	50	.72**	.49**	26	.63**	0.07	23	.83	0.63**			
6	47	.69**	.46**	23	.68**	0.49**				24	.68**	.48*

* $p < .05$. ** $p < .01$. *** $p < .001$.

Validity

Two types of analysis involving evidence based on relations with other variables, specifically convergent-discriminant evidence and concurrent evidence, were conducted with the LEO-1-min scores.

Convergent-discriminant evidence about the LEO-1-min as a task to measure word-reading performance was examined by calculating Pearson's correlations between the scores on the LEO-1-min Form A and the scores on the word and pseudoword subtests of the PROLEC-R and the Arithmetic test. Table 3 shows the high correlations achieved in all grades between the performance on the form A of the LEO-1-min and the PROLEC-R, especially for word reading. By contrast, the correlations of the LEO-1-min Form A with the scores on the Arithmetic were lower, and most of them were non-significant, in all grades except for 1st and 2nd. It should be noted that 4th grade showed lower correlations with the PROLEC-R and with Arithmetic. This result is possibly due to the low variability in the LEO-1-min scores in this grade.

Table 3. *Pearson's Correlations among the Correct Responses on the LEO-1-min Form A, on PROLEC-R and on Arithmetic by Grade.*

	PROLEC-R		Arithmetic
	Words	Pseudowords	
LEO-1-min Form A			
1 (*n*= 17)	.60*	.51*	.56*
2 (*n*= 18)	.74**	.60**	.59**
3 (*n*= 16)	.85**	.82**	.44
4 (*n*= 17)	.55*	.38	.24
5 (*n*= 14)	.72**	.70**	.38
6 (*n*= 17)	.58*	.58*	.22

* $p < .05$. ** $p < .01$. *** $p < .001$.

Equivalence is another index of validity. If several forms of a test measure the same thing, these forms should correlate with a demographic variable to a similar degree, compared to one another. Table 4 shows the magnitudes with which the LEO-1-min forms correlate with age and gender. All four forms correlate moderately with age; by contrast, correlations with gender are modest.

Predictive discriminant analysis was used to examine to what extent the test score predicted the reading level assigned by the teachers. The predictive variable was the score on the LEO-1-min Form A, and the grouping variable was the reading level. To maximize the difference in reading levels, we grouped together children from normal and high levels. In this way, the normal-high performance group was made up of anyone who received scores of 1 and 2 from their teachers, and the low performance group consisted of those who had received a score of 3 from their teachers. The normality of the predictive variable was fulfilled according to the Kolmogorov test (p = 0,064), and Box's test for the equality of the variance-covariance matrix was not significant. $F < 1$. The discriminant analysis revealed a significant discriminant function, χ^2 (1, N = 102) = 18,747, $p < .000$ (eigenvalue = .207, canonical correlation = .414, Wilks's Λ= .828). The mean of the centroid groups in the discriminant function revealed that the normal-high reading group achieved a positive mean (.347), and the low-level group achieved a negative mean (-.585). These results show that the normal-high performance group read a higher number of words and identified a higher number of pseudowords on the LEO-1-

min than the low performance group. Implementing the Leave-One-Out method to control the tendency to overestimate the accuracy of the classification rates, the correct classification rate for the whole sample was 72%. Huberty's Z indicated that the rate of correct classifications was statistically better than what would be expected at random, $z = 2.02$, $p = .002$.

Norms. To complete the creation of the LEO-1-min, we elaborated standards with the study sample (1st to 6th grades) that allowed us to transform the direct scores into centiles (see Table 5). This makes it possible to place the children within their normative reference group. Note that an unintended result is that the quartiles of the LEO-1-min match the mean scores that described the three reading-level groups according to the teachers.

Table 4. *Pearson's Correlations among Age, Sex, and Forms of LEO-1-min.*

	LEO-1-min			
	Form A	Form B	Form C	Form D
	n = 283	n = 146	n = 67	n = 69
Age	.765**	.729**	.805**	.725**
Sex[a]	-.017	.002	.038	-.063

[a]Due to the missing data for sex, *n* for Forms A, B, C, and D are 155, 64, 32, and 59, respectively.

** $p < .01$.

Discussion and conclusion

The study illustrates the potential usefulness of the LEO-1-min to measure word reading in a wide age range, even though here we only discuss the evidence obtained from a sample of primary school children. The performance on the test showed the progressive development of reading throughout primary education, observed in previous studies in Spanish (Goikoetxea & Ferrero, 2019). Children from first and second grade make accelerated progress, with slower growth from 3rd grade, which makes Grades 3 and 4 and Grades 5 and 6 very similar to each other. It is interesting to again observe the low number of errors even from Grade 1, as observed in Spanish (Valle-Arroyo, 1989) and other transparent orthographies (Landerl, & Wimmer, 2008), compared to opaque ones (Seymour, Aro, Erskine, & COST Action A8, 2003; Simöes & Alves, 2018). Unlike Van Bon (2007), who found a signif-

icant but small (and societally probably irrelevant) difference, we did not find any effect of sex on the performance on the LEO-1-min. However, the sex data loss in our study made this result tentative. Also in contrast to Van Bon (2007), the performance on the LEO-1-min correlates with the performance on the symbols task. A tentative explanation for these results is that Van Bon (2007) used Greek letters on the symbols task that are completely meaningless to the readers of our alphabets. But in this study we used letter strings, and from Stroop-like tasks, we know that even nonsense letter strings distract the attention and delay the naming of pictures because competent readers cannot keep themselves from reading.

Regarding the equivalence of the LEO-1-min forms, Forms A and B demonstrate similar mean scores, and the distribution of scores is also quite similar across both forms. Due to the high level of comparability between forms A and B, these two forms may be considered equivalent in difficulty. Forms C and D also reveal similar means and standard deviations to those of form A, at least in the three grades compared. In addition, the magnitude of the retest reliability coefficients is moderate to high in all grades, reaching the minimum requirements for a test to be used in decision making in samples similar to the one in this study. The reliability coefficients are similar to those obtained in the Netherlands with 2nd and 3rd grade children (Van Bon et al., 2004). These results suggest that forms A and B can be considered equivalents forms, making them useful for repeated examination throughout primary school. Although we do not examine the order effect and, consequently, we do not know the effect of practice on the LEO-1-min, the similarity in the means suggests that there is no order effect. However, future research must address this question.

Regarding convergent-discriminant validity evidence, it can be observed that the scores on oral reading on the LEO-1-min strongly correlate with the PROLEC-R, and to a lesser extent with the Arithmetic test. The correlation between lexical decision and reading aloud supports the validity of the lexical-decision task as a measure of the cognitive processes in word reading. The lower correlation with Arithmetic shows that our test does not indicate general school achievement, but rather reading skills in particular. Furthermore, the LEO-1-min scores discriminate between groups of good and poor readers identified by the teachers. This evidence of the concurrent validity of the test is an indicator of its possible usefulness in performing activities that increase word reading skills and reading fluency.

Despite the small and accidental sample in this study, provisional scales are presented based on our data. The scales can be applied in the school context and in research on reading performance, but with caution because our sample comes from only one school in the province of Vizcaya.

The obvious limitations of this study, warranting further investigation, are the fact that we have not provided performance data from testing a representative sampling of the primary school population, or stability data for the measures (test-retest), or validity tests of the scores over time. Likewise, it is necessary to know the psychometric properties of the LEO -1-min in representative samples to evaluate their use in Spanish youth and adult populations. The data offered here cannot be generalized beyond the characteristics of the participants in this study.

In conclusion, the LEO-1-min scores show reliability and validity as a measure of word reading skills in Spanish-speaking children with the characteristics of the sample participating in this study. Good teaching of reading requires repeated assessments of the same individuals over time. The LEO-1-min will allow teachers and researchers to perform a quick, groupwise, and repeated assessment of children's word-reading skills. Today we know that the early detection of low reading performance throughout the primary grades is the best remediation intervention because it enables teachers to react immediately, adapting their reading instruction to individual needs.

References

Acha, J., & Perea, M. (2008). The effects of length and transposed-letter similarity in lexical decision: Evidence with beginning, intermediate, and adult readers. *British Journal of Psychology, 99* (2), 245-264. doi:10.1348/000712607x224478

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: Authors.

Carreiras, M., Alvarez, C.J., & de Vega, M. (1993). Syllable frequency and visual word recognition in Spanish. *Journal of Memory and Language, 32*, 766-780. doi:10.1006/jmla.1993.1038

Corral, S., Ferrero, M., & Goikoetxea, E. (2009). LEXIN: A lexical database from Spanish kindergarten and first-grade readers. *Behaviour Research Methods*, *41*(4), 1009-1017. doi:10.3758/brm.41.4.1009

Cuetos, F., Rodríguez, B., Ruano, E., & Arribas, D. (2007). *PROLEC-R: Battery of evaluation of reading processes for children from primary education reviewed.* Madrid: TEA.

Davis, C. J., & Perea, M. (2005). BuscaPalabras: A program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indexes in Spanish. *Behavior Research Methods, 37*, 665-671. doi:10.3758/bf03192738

Defior, S., Fonseca, L., Gottheil, B., Aldrey, A., Rosa, G., Pujals, M., … Serrano, F. D. (2006). *LEE. Test de lectura y escritura en español.* Buenos Aires: Paidós.

Defior, S., Justicia, F., & Martos, F. (1996). The influence of lexical and sub lexical variables in normal and poor Spanish readers, *Reading and Writing, 8*, 487-497. doi:10.1007/bf00577024

Forster, K. I. (1976). Accessing the mental lexicon. In R. J. Wales & E. Walker (Eds.), *New approaches to language mechanisms* (pp. 257-287). Amsterdam: North-Holland.

Forster, K. I., & Chambers, I. M. (1973). Lexical access and naming time. *Journal of Verbal Learning & Verbal Behavior*, *12*, 627–635. doi:10.1016/s0022-5371(73)80042-8

Foster, T. E., Ardoin, S. P., & Binder, K. S. (2018). Reliability and validity of eye movement measures of children's reading. *Reading Research Quarterly, 53 (1),* 71-89. doi:10.1002/rrq.182

Förster, N., Kawohl, E., & Souvignier, E. (2018). Short-and long-term effects of assessment-based differentiated reading instruction in general education on reading fluency and reading comprehension. *Learning and Instruction, 56,* 98-109. doi.org/10.1016/j.learninstruc.2018.04.009

Försters, N., & Souvignier, E. (2015). Effects of providing teachers with information about their students' reading progress. *School Psychology Review, 44 (1),* 60-75. doi:10.17105/spr44-1.60-75

Goikoetxea, E., & Ferrero, M. (2019). Word recognition growth in Spanish children. Manuscript submitted for publication.

González-Trujillo, M. C., Calet, N., Defior, S., & Gutiérrez-Palma, N. (2014) Escala de fluidez lectora en español: midiendo los componentes de la fluidez, *Estudios de Psicología, 35 (1),* 104-136. doi:10.1080/02109395.2014.893651

Jenkins, J. R, Fuchs, L. S., van den Broek P, Espin C, & Deno S. L. (2003). Sources of individual differences in reading comprehension and reading fluency. *Journal of Educational Psychology, 95,* 719–729. doi:10.1037/0022-0663.95.4.719

Jiménez, J. E., Gove, A., Crouch, L., & Rodríguez, C. (2014). Internal structure and standardized scores of the Spanish adaptation of the EGRA (Early Grade Reading Assessment) for early reading assessment. *Psicothema, 26(4),* 531-537.

Keenan, J. M., & Meenan, C. E. (2014). Test differences in diagnosing reading comprehension deficits. *Journal of Learning Disabilities, 47 (2),* 125-135. doi:10.1177/0022219412439326

Kim, Y.S., Petscher, Y., Schatschneider, C., & Foorman, B. (2010). Does growth rate in oral reading fluency matter in predicting reading comprehension achievement? *Journal of Educational Psychology, 102,* 652–667. doi:10.1037/a0019643

Klauda, S. L., & Guthrie, J. T. (2008). Relationships of three components of reading fluency to reading comprehension. *Journal of Educational Psychology, 100 (2),* 310-321. doi:10.1037/0022-0663.100.2.310

Kuperman, V., Drieghe, D., Keuleers, E., & Brysbaert, M. (2013). How strongly do word reading times and lexical decision times correlate? Combining data from eye movement corpora and megastudies. *The Quarterly Journal of Experimental Psychology, 66 (3),* 563–580. doi.org/10.1080/17470218.2012.658820

Lacina, J., & Block, C. (2011). What matters most in distinguished literacy teacher education programs? *Journal of Literacy Researc*h, 43(4), 319-351. doi:10.1177/1086296x11422033

Landerl, K., & Wimmer, H. (2008). Development of word reading fluency and spelling in a consistent orthography: An 8-year follow-up. *Journal of Educational Psychology, 100(1),* 150-161. doi:10.1037/0022-0663.100.1.150

Marín, J., & Carrillo, M. S. (1999). *Test Colectivo de Eficacia Lectora (TECLE).* Unpublished manuscript, Universidad de Murcia.

Moret-Tatay, C., & Perea, M. (2011). Is the go/no-go lexical decision task preferable to the yes/no task with developing readers. *Journal of Experimental Child Psychology, 110 (1),* 125-132. doi:10.1016/j.jecp.2011.04.005

Perea, M., & Rosa, E. (2000). The effects of orthographic neighborhood in reading and laboratory word identification tasks: A review. *Psicológica, 21,* 327-340.

Perfetti, C. & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading, 18(1),* 22-37. doi:10.1080/10888438.2013.827687

Rasinski, T. V., Reutzel, C. R., Chard, D. & Linan-Thompson, S. (2011). Reading fluency. In M. L. Kamil, P. D. Pearson, B. Moje, & P. Afflerbach E. (Eds), *Handbook of Reading Research, Volume IV* (pp. 286-319). New York: Routledge.

Rayner, K., & Reichle, E. D. (2010). Models of the reading process. *Wiley Interdisciplinary Reviews: Cognitive Science, 1 (6)*, 787-799. doi:10.1002/wcs.68

Rayner, K., Sereno, S., Morris, R., Schmauder, R., & Clifton, C., Jr. (1989). Eye movements and on-line language comprehension processes. *Language and Cognitive Processes, 4*, 21–49. doi:10.1080/01690968908406362

Schilling, H. E. H., Rayner, K., & Chumbley, J. I. (1998). Comparing naming, lexical decision, and eye fixation times: word frequency effects and individual differences. *Memory & Cognition, 26 (6)*, 1270-1281. doi:10.3758/bf03201199

Sebastián-Gallés, N., Martí, M. A., Carreiras, M. F., & Cuetos, F. (2000). *LEXESP: Léxico informatizado del español* [LEXESP: A computarizad word-pool in Spanish]. Barcelona: Universitat de Barcelona.

Seymour, P. H., Aro, M., Erskine, J. M., & COST Action A8 (2003). Foundation literacy acquisition in European orthographies. *Bristish Journal of Psychology, 94*, 143-174. doi:10.1348/000712603321661859

Simöes, E., & Alves-Martins, M. (2018). Reading acquisition in beginner readers: Typical errors in European Portuguese. *Educação e Pesquisa*, *44*, e165734. Retrieved from http://dx.doi.org/10.1590/S1678-4634201844165734

Valle-Arroyo, F. (1989). Reading errors in Spanish. In P. G. Aaron & R. M. Joshi (Eds.), *Reading and writing disorders in different orthographic systems* (pp. 163–175). The Netherlands: Kluwer Academic.

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

Van Bon, W. H. J., Hoevenaars, L. T. M., & Jongeneelen, J. J. (2004). Using paper-and-pencil lexical-decision tests to assess word decoding skills: Aspects of validity and reliability. *Journal of Research in Reading, 27 (1)*, 58–68. doi:10.1111/j.1467-9817.2004.00214.x

Van Bon, W. H. J., & Libert, J. E. A. (1997). Oral reading and silent reading compared: Evidence for a subtype of poor readers. *Polish Psychological Bulletin, 28,* 59-70.

Van Bon, W. H., Tooren, P. H., & van Eekelen, K. W. (2000). Lexical decision and oral reading by poor and normal readers. *European Journal of Psychology of Education, 15(3)*, 259-270. doi:10.1007/bf03173178

Wechsler, D. (2005). *Escala de inteligencia de Wechsler para niños IV*. Madrid: TEA.

Wilson, J. M. G., & Jungner, G. (1968). *Principles and practice of screening for diseases*. Geneva: World Health Organization.

Appendix A

LEO-1-min

TEST LEO-1-min

Name:

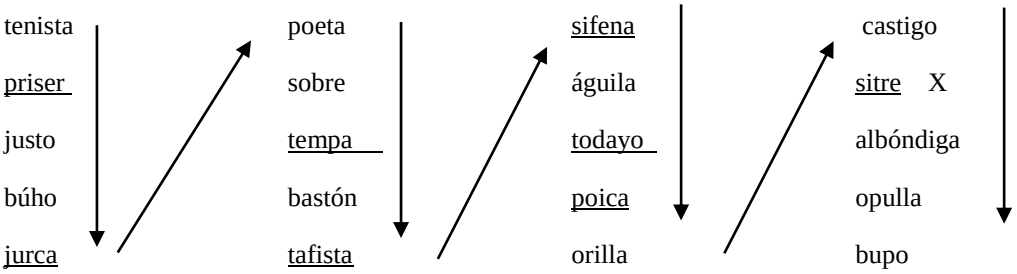
Age:

Grade:

INSTRUCTIONS

1. You are going to find some words written in Spanish in several columns.
2. Some of those words truly exist, and others are invented.
3. Underline or draw a line underneath the words that do not exist.
4. Start with the column on the left from top to bottom.
5. If you finish the first column, go on to the next one on the right-hand side. Always follow this order.
6. When the time is up you will hear: "STOP". At that moment, stop reading and put an X to the right of the last word, real or invented that you have read.

Example:



Appendix B

LEO-1-min Form A

amigo	tarea	cabeza	sopefad
pretio	sur	papel	doctor
fuego	blanco	atención	piso
ayuta	literatura	gato	daño
pol	piano	par	barfe
imagen	río	encuentro	valor
morenso	presente	contenido	objeto
conjunto	carne	reamión	codi
lista	vicina	capital	loño
flus	patio	relisma	ayer
media	resultado	vino	francés
sala	cine	título	motor
rul	kilómetros	signo	alcohol
payamo	semana	borde	tipo
vía	secunto	unidad	prejitente
oportunidad	díe	letra	plan
red	abuelo	teléfono	caja
cazano	prenfa	cola	ciendia
resto	cuarenta	caso	sombra
radio	sueño	imborcancia	lado
trel	época	hermano	gloria
casirad	serie	humor	rojo
habitación	catanidad	color	estado

Appendix B (continuation)

hambre	madera	curso	cuarlo
pena	rato	aspecto	dueño
espajo	esbrimor	pelo	cocina
esfuerzo	inteligencia	baru	café
dama	seguridad	ripo	centro
frase	nada	pie	intención
televisión	cantidad	policía	plaro
planeta	programa	nombre	piedra
cultura	peligro	otode	sitio
bope	impresión	sentimiento	jardín
ojo	bar	tierra	nitel
gente	mado	canal	guena
escena	ciepa	luba	alión
vaso	producto	pleno	favor
hujo	revolución	vida	fatibia
deseo	poatía	trabajo	palabra
quinle	energía	máquina	medida
estación	mundo	llegada	fiesca
renueldo	decisión	lunes	problema
frecuencia	puerta	carta	
cona	comida	forsagión	
cura	cacilán	mayoría	
espectáculo	salud	térjilo	

		Response	
		Item underlined	No action
Item	Pseudoword	Hit	Omission
	Word	False alarm	Correct omission

Figure 1. *Matrix stimulus-response*

Appendix C

Table 5. *Provisional Norms from LEO-1-min Form A in Centiles per Grade*

	Grade					
	1	2	3	4	5	6
Centile	*n* = 48	*n* = 47	*n* = 51	*n* = 42	*n* = 50	*n* = 46
95	41	48	62	76	91	97
90	31	41	55	63	78	85
80	21	36	52	54	72	77
70	16	34	51	53	67	69
60	14	30	45	50	64	66
50	13	26	44	48	60	65
40	11	24	36	45	56	55
30	10	21	32	42	53	53
20	6	19	31	39	48	49
10	3	16	21	34	40	40
5	2	15	20	26	33	36
Mean	15.17	28.74	41.53	48.17	61.62	63.24
(*St*)	(10.75)	(11.59)	(13.09)	(11.71)	(17.55)	(15.71)

Recibido: 08-10-2018
Aceptado: 04-04-2019