

# A letter visual-similarity matrix for Latin-based alphabets

Ian C. Simpson · Petroula Mousikou ·  
Juan Manuel Montoya · Sylvia Defior

Published online: 6 October 2012  
© Psychonomic Society, Inc. 2012

**Abstract** Indicators of letter visual similarity have been used for controlling the design of empirical and neuropsychological studies and for rigorously determining the factors that underlie reading ability and literacy acquisition. Additionally, these letter similarity/confusability matrices have been useful for studies examining more general aspects of human cognition, such as perception. Despite many letter visual-similarity matrices being available, they all have two serious limitations if they are to be used by researchers in the reading domain: (1) They have been constructed using atypical reading data obtained from speeded reading-aloud tasks and/or under degraded presentation conditions; (2) they only include letters from the English alphabet. Although some letter visual-similarity matrices have been constructed using data gathered from normal reading conditions, these either are based on old fonts, which may not resemble the letters found in modern print, or were never published. For the first time, this article presents a comprehensive letter visual-similarity/confusability matrix that has been constructed based on untimed responses to clearly presented upper- and lowercase letters that are present in many languages that use Latin-based alphabets, including Catalan, Dutch, English, French, Galician, German, Italian, Portuguese, and Spanish. Such a matrix will be useful for

researchers interested in the processes underpinning reading and literacy acquisition.

**Keywords** Letter visual-similarity/confusability matrix · Latin alphabets · Reading · Literacy acquisition

In an extensive review of the literature spanning over 120 years, Mueller and Weidemann (2012) identified more than 70 published articles that had in one way or another sought to measure letter similarity (or confusability). From this review, the authors identified three main motivations for studying letter similarity: (1) practical attempts to make written text more comprehensible, and thereby allow learners to acquire reading skill more easily; (2) empirical investigations with the goal of understanding how the visual system functions; and (3) theoretical research attempting to explain how letters are represented by the visual system or, in abstract form, by the cognitive system. Regardless of the motivations for studying letter similarity, in the majority of these studies the same basic paradigm has been used—that of presenting single letters to participants, with the task being to name the presented item. A confusion matrix is then constructed by noting how often each letter was (incorrectly) given as a response to the presented letter. The number of responses given for each stimulus–response letter pair has been assumed to be an indication of the level of similarity (or confusability) between the two letters, with more errors on a pair indicating higher similarity/confusability. One problem with this paradigm is that if the letters are presented in a visually clear manner and no time limit to make a response is imposed on the participants, few errors are generated, and therefore no letter similarity/confusability matrix can be constructed. As a result, many variations have been introduced to this paradigm so that errors will occur frequently. These include using brief presentations of the letters (Mueller & Weidemann, 2012), reducing the letter

---

**Electronic supplementary material** The online version of this article (doi:10.3758/s13428-012-0271-4) contains supplementary material, which is available to authorized users.

---

I. C. Simpson (✉) · J. M. Montoya · S. Defior  
Departamento de Psicología Evolutiva y de la Educación,  
Universidad de Granada,  
Granada, Spain  
e-mail: isimpson@ugr.es

P. Mousikou  
School of Psychology, Bangor University,  
Bangor, North Wales, UK

size (Phillips, Johnson, & Browne, 1983), presenting letters in the peripheral visual field (Reich & Bedell, 2000), and creating low-contrast or noisy viewing conditions (Liu & Arditi, 2001). While all of these techniques do increase the number of errors generated by participants, several problems are associated with them if letter visual-similarity matrices constructed according to these techniques are to be used by researchers in the reading domain.

For example, Fiset et al. (2009) argued that demanding experimental manipulations such as low contrast or rapid presentation exacerbate the relative importance of visual information that is not optimal for human vision, thus leading to very high error rates. Accordingly, Fiset et al. asserted that these paradigms may be inadequate for the discovery of the letter features underlying reading in daily life. This seriously limits the use of such matrices among reading researchers who aim to understand how letter knowledge, a cornerstone for successful reading, is acquired. Furthermore, participants may commit confusion errors in this sort of paradigm for several reasons, with visual similarity between the target and response being only one possible factor. In speeded naming conditions, for example, the phonological similarity of the target's letter name to the response's letter name (e.g., "ef" and "ess"), or between the phonemes usually represented by each letter (e.g., /b/ and /p/), may also be contributing factors (Treiman & Kessler, 2004). Accordingly, existing letter similarity matrices that have been constructed on the basis of findings from this basic paradigm may represent the overall confusability between letter pairs, but nevertheless, they do not allow for individual dimensions of similarity, such as visual similarity, to be isolated from other similarity dimensions, such as phonological similarity. This seriously limits their use as a tool for well-controlled psycholinguistic studies that aim to uncover the processes underlying our ability to acquire and master reading.

An additional limitation of all of the similarity matrices published to date is that, except for a single similarity matrix containing the lowercase Swedish letters "å", "ä", and "ö" (Kuennapas & Janson, 1969), all others have been constructed using only the 26 letters of the English alphabet. This means that no published data are available for letters such as "Ñ" from Spanish or "ß" from German, which have their own names and are taught as wholly separate letters of the alphabet. Similarly, there are no available data for letters with diacritic marks. While these forms are rarely considered to be separate letters and usually do not form part of the alphabet, the presence of diacritic marks often changes stress assignment, pronunciation, or the meaning of a word. To be considered a competent reader in a language that includes diacritic marks, it is therefore important to recognize the difference between letters that do and do not carry these marks.

To illustrate why it is important to have a letter visual-similarity matrix that extends beyond the 26 letters of the English alphabet, we now consider the example of letter knowledge—a topic of interest to researchers investigating how children learn to read in different alphabetic orthographies. A simple way to assess a young child's letter knowledge is to ask the child to provide the letter name or letter sound to visually presented single letters. Among the various types of errors that children make in such tasks are letter confusion errors—that is, providing the name or sound of a letter different from the one presented (e.g., responding with the name of "b" when presented with the letter "d"). Such errors are commonly observed in many alphabetic languages (e.g., for English and Portuguese, see Treiman, Kessler, & Pollo, 2006; for Spanish, see Goikoetxea, 2006). Identifying the factors that drive letter confusion errors is critical for understanding how children acquire letter knowledge, which is known to be one of the best predictors of later reading ability (Snow, Burns, & Griffin, 1998). Given that Treiman, Levin, and Kessler (2007) demonstrated that letter visual similarity can contribute to letter confusion errors independently of letter phonological similarity, the lack of a letter visual-similarity matrix that contains symbols from different alphabetic languages poses another serious limitation for researchers who are interested in understanding the processes underlying reading acquisition in languages other than English. The availability of such a matrix would also facilitate cross-linguistic research in this domain.

It is worth noting that the utility of a letter visual-similarity matrix is not limited to research related to language acquisition; yet, in setting out to construct such a matrix, it was important to ensure that it would also be applicable to this area of investigation. As children normally read under good viewing conditions and with no time pressure, it would seem most appropriate to analyze the visual aspect of child confusion errors using a visual similarity matrix constructed from data gathered using natural reading conditions. We are aware of only one study published within the last 25 years that presents a letter visual-similarity matrix gathered using a methodology of normal presentation—that of Boles and Clifford (1989). However, several problems exist with this matrix. Firstly, the font used was Apple-Psych, an old computer-based font with letters constructed using a 6 × 7 dot matrix. By today's standards, this style of font is considered coarse, and the letters themselves appear somewhat compressed or squat. Therefore, the similarity ratings for some combinations of letters in the Boles and Clifford matrix may not be accurate representations of the similarity between the letters when they are printed in a more modern font.

Secondly, Boles and Clifford (1989) included pairs of identical letters among their stimuli (e.g., rating the visual

similarity of “A” and “A”). This is not in itself a problem. However, if one inspects the ratings given to the identical pairs, it becomes apparent that for a number of these pairs the mean similarity ratings are less than perfect. For example, the matching pair “u u” (mean rating 4.42 out of a possible 5.00) was considered less similar than the non-matching pair “q g” (mean rating 4.67). Similarly, while the matching pair “k k” was considered a perfect match (mean rating of 5.00), the matching pair “K K” was considered less than a perfect match (mean rating 4.75). On the surface, this might indicate that the participants truly thought that the pair “K K” was somehow less of a match than the pair “k k.” However, these results may well indicate the participants’ fatigue in performing this task. Each of the 32 participants in Boles and Clifford’s study was required to rate over 1,000 pairs of letters. Yet our own pilot testing indicated that after approximately 200 trials, participants became less sensitive to the visual similarities of the stimuli, in the sense that they assigned, unreasonably, the lowest possible rating (*not at all similar*) more often than they did in the initial trials. In the Boles and Clifford matrix, more than one quarter of the matching pairs received a mean rating of less than 5, suggesting that potentially 25 % of the ratings contained in their data set were noisy or adversely influenced by uncontrolled factors.

The shortcomings of the available letter visual-similarity matrices for use in analyzing reading data are demonstrated by the fact that recent studies that have analyzed the visual similarity of child letter confusion errors in different languages (e.g., English, Portuguese, and Hebrew) have constructed their own similarity matrices (Treiman et al., 2006; Treiman, Levin, and Kessler 2007). In each case, the method used to construct the rating of visual similarity was the same. Firstly, all of the letter confusion errors contained in the child data were identified. The full set of letter confusion errors was then presented to 30 undergraduates, who were asked to rate the letter pairs on the basis of visual similarity. By taking the mean of the responses, a similarity scale was constructed that could be used to analyze the data in each study. Besides being effortful to construct, these letter similarity data were collected from undergraduates with the purpose of analyzing the specific set of letter confusion errors found within the child data. As a result, such matrices did not contain ratings for all possible letter combinations, and therefore were never published. This means that subsequent investigators wishing to perform similar research would need to repeat the process of constructing their own letter visual-similarity matrix. Even if the previous researchers were to make their matrices available, the data would contain only a subset of all possible letter pairs. The main aim of the present article was thus to provide a comprehensive letter visual-similarity matrix that could be used in this area of research in the future.

The need for such a matrix in the literacy acquisition, and more generally in the reading, domain is evident from the fact that the letter visual-similarity matrix of Boles and Clifford (1989), which was constructed using the paradigm most relevant for reading research—presenting the stimuli clearly and in an untimed manner—has been cited by approximately 20 studies in this domain in the last 6 years. Within these studies, the Boles and Clifford matrix has been used in a number of ways. For example, Burgund and Abernathy (2008) examined the extent to which children and adults are sensitive to the visual forms of letters when reading, and whether such sensitivity depends on reading skill. To construct their stimulus materials, they controlled for the visual similarity between letters using the Boles and Clifford matrix. Fiset, Gosselin, Blais, and Arguin (2006) used the same matrix to explain aspects of pure alexia, a reading disorder that is characterized by abnormal sensitivity to letter confusability. Treiman, Cohen, Mulqueeny, Kessler, and Schechtman (2007), on the other hand, used the Boles and Clifford matrix to investigate the nature and development of young children’s knowledge about writing.

To summarize, although many visual similarity matrices for letters are presently available, all but one of the matrices published in the last 120 years are limited to the 26 letters of the English alphabet. In addition, in almost all cases the matrices were based on data generated in atypical reading conditions. Such matrices are therefore unsuitable for researchers in the reading domain or for those interested in carrying out cross-linguistic studies in this field. Our goal in this article was to address these points by creating a visual-similarity matrix for letters that could mainly be used by investigators carrying out research in reading in different alphabetic languages. The technique that we employed to construct the similarity matrix was the same as the one used in previous studies examining letter knowledge in children (Treiman et al., 2006; Treiman, Levin, and Kessler 2007) and is based on data obtained under normal (untimed) reading conditions. We anticipate that the matrix presented here will also prove useful to researchers in any field of investigation in which Latin letters are used as stimuli and a measure of visual similarity between stimuli is required.

In the first study, the participants consisted of speakers of Spanish who were required to rate letter pairs using a scale of 1 (*not similar at all*) to 7 (*very similar*). The letter pairs were formed using letters included in many Latin-based alphabets. In the second study, a different group of Spanish participants were similarly asked to rate letter pairs that were formed using all of the letters of the Spanish alphabet, bar the ones containing diacritic marks. The second study was carried out to test whether the inclusion of diacritic marks in the first study had influenced participants’ visual similarity ratings of letter pairs that contained no diacritic marks. In the third study, the participants were speakers of English

who were asked to rate, in a similar manner, letter pairs that were formed using all of the letters of the English alphabet. The latter study was carried out to test whether participants' visual similarity ratings of letter pairs were influenced by cultural and/or linguistic factors.

## Method

### Participants

For the first study, 677 participants were recruited for the initial data collection round. An additional 55 participants were recruited for a second data collection round to replace participants from the first round whose data were excluded from the analysis (see the Results section for details). All 732 participants were undergraduate students in either the Faculty of Psychology or the Faculty of Education Science at the University of Granada, Spain. Of these, six were nonnative speakers of Spanish, although all of the nonnatives were nevertheless familiar with Latin-based alphabets (three Germans, two French, and one Portuguese). Recruitment and participation took place during the students' normal classes. Participation was voluntary, with some students opting not to take part. The students received neither payment nor class credit for their participation.

### Selection of stimuli

We selected letters from a variety of Latin-based alphabets, including Dutch, English, French, German, Italian, Portuguese, and Spanish. These particular languages are the most widely studied in psycholinguistic research, and therefore a comprehensive letter visual-similarity matrix that includes all of the letters used in these languages may prove useful for advancing research in this area. The visual-similarity matrix, which contains a list of all of the stimuli included in the study, is available in the [online supplementary materials](#).

Apart from identifying specific letters that appeared in the languages that we wanted to cover, children's reading material presents an additional complexity. Given that our letter visual-similarity matrix is addressed to researchers interested in developmental reading processes, it was important to consider the types of fonts that children are most likely to encounter. Most fonts used in reading materials designed for children are sans serif (Walker & Reynolds, 2003; Wilkins, Cleave, Grayson & Wilson, 2009). In particular, the sans serif font Arial has been shown to be popular among both children (Bernard, Chaparro, Mills, & Halcomb, 2002) and adults (Bernard, Chaparro, Mills, & Halcomb, 2003); therefore, we decided to have the majority of letters in our study displayed in this font. However, the shapes of

the characters used for children may differ from those used for adults—particularly with regard to the single-story lowercase “c” and “g” (known as *infant form*; Walker & Reynolds, 2003) versus their two-story counterparts “a” and “g” (known as *adult form*). Walker and Reynolds have argued that the adult forms of the letters “a” and “g” are familiar to children who have begun to read and may be preferred to the infant forms because they are less confusable. Despite this, it is widely assumed that the infant forms of these letters are easier for children (Walker & Thiessen, 2011), and the infant forms are most commonly used in reading materials directed at children (Walker & Reynolds, 2003). For this reason, when considering which form of the letters “a” and “g” to use (which, in the case of the letter “a”, meant six different variations—“ɑ”, “ä”, “á”, “â”, “ã”, and “ä”), the infant forms were chosen. Thus, for all variations of lowercase “a”, this meant that a different font was required, and Comic Sans was chosen, as it has been shown to be popular with children (Bernard et al., 2002; Taslim, Wan Adnan, & Abu Bakar, 2009). Given that the Boles and Clifford (1989) matrix contained the adult form of lowercase “a” and a serif form of uppercase “I”, for comparison reasons, it was decided to include these alternate forms as additional items in our matrix. In the case of “I”, this meant choosing a serif font, so Times New Roman was chosen.

To summarize, as we envisage that the present matrix will be most widely used in the domain of literacy acquisition, we wanted to use letters that were commonly found in children's reading materials. This meant presenting letters in a sans serif font, and also presenting the lowercase letters “c” and “g” in the infant form, which is commonly used in children's reading materials. Furthermore, in the case of lowercase “a” and uppercase “I”, alternative forms of these letters were included in the stimuli. This was done to allow for comparison of the present matrix with that of Boles and Clifford (1989), which is the matrix that is most relevant to reading research that has been published to date. In total, we presented 52 uppercase letters<sup>1</sup> and 53 lowercase letters—the difference being due to the fact that the German letter “ß” exists only in lowercase form. Combining the uppercase letters resulted in a total of 1,326 unordered pairings,<sup>2</sup> while combining the lowercase letters produced 1,378 unordered pairings (see the [visual-similarity matrix in the online supplementary materials](#)).

<sup>1</sup> For the purposes of this article, we will use the term “letter” to refer to any stimulus that is visually different from any other stimulus (e.g., “E” and “É”), even if the stimulus is not considered to be a letter in its own right in any alphabet.

<sup>2</sup> Unlike Boles and Clifford (1989), we did not include identical pairs, such as “A A”, in our stimuli.

## Apparatus and materials

Response booklets consisting of one page of instructions followed by ten pages of stimuli were given to the participants. Printed at the top of each stimulus page was the rating scale that participants were to follow, and underneath this, two columns of six letter pairs were printed. The letters in each pair were separated by a single space with an area provided next to each of the 12 letter pairs for the participants' responses. Thus, each response booklet contained 120 pairings. The 120 letter pairings in each booklet were selected at random from the complete list of pairings, although each booklet contained entirely upper- or entirely lowercase letters. Within each response booklet, letter pairings were randomly ordered, with the restriction that the same letter did not appear in two consecutive pairings. As each letter pair appeared in multiple booklets, the ordering of the items within each pair was counter-balanced across response booklets, with half occurring in one order (e.g., "A B") and the other half in the alternative order (e.g., "B A"). All letters were presented in Arial 72-point (a sans serif) font, except for (1) the second uppercase version of the letter "I", which was presented in Times New Roman 72-point font, so that it was presented with serifs, and (2) the second lowercase "c" and the five forms of lowercase "c" with diacritic marks ("ç", "ć", "ĉ", "ċ", and "ö"), which were presented in Comic Sans 72-point font so that they were displayed in the infant form of lowercase "c".

## Design and procedure

Of the initial 677 participants, 332 were randomly assigned to rate uppercase letter pairs, with the remaining 345 participants rating the lowercase letter pairs. As all of the participants rated exactly 120 pairs, the majority of pairings were rated by 30 participants, although 60 of the 1,326 uppercase pairings and 60 of the 1,378 lowercase letter pairings were rated by 31 participants. Fifty-five extra students were recruited to replace participants who were excluded (as described below). In this instance, the new participants received the same letter pairs as had the excluded participants who they replaced. Participants were instructed to ignore the sounds of the letters and to rate the letter pairs purely on visual similarity, using a scale from 1 (*not at all similar*) to 7 (*very similar*). Participants were also warned that some of the letters might be unfamiliar to them, but they were reminded that, since they were only rating visual similarity, familiarity should not affect the ratings given. No time limits were imposed and participants responded at their own pace, with all participants completing the task in less than 20 min.

## Results and discussion

The written responses were transferred to digital form, where an initial screening process was performed in order to detect obvious cases in which the participants may have misunderstood or not correctly followed the instructions. This resulted in the data from 23 participants being excluded from further analyses, for the following reasons: most or all responses blank (11 participants), only ratings of 1 or 7 used (four participants), only ratings of 1 or 2 used (two participants), probable reverse interpretation of the scale (e.g., assigning a value of 1 to pair "A Á" and a value of 7 to pair "E Ç"; two participants), assigning responses using values outside of the specified range (e.g., assigning ratings of 0 and 10; one participant), and obvious random responses (three participants).

To detect less obvious cases of incorrect responding, we identified participants whose patterns of responses appeared to differ from those of the overall sample. This was done with a two-step process. Firstly, the mean rating and standard deviation were calculated for each of the 1,326 uppercase and 1,378 lowercase letter pairs. Secondly, for each participant, the number of individual responses that fell outside of a range of  $\pm 2$  SDs of these means was calculated. Any participants with more than 25 % of their responses falling outside of this range were excluded. A further 32 participants were excluded using this criterion. In all, the data from 55 participants were excluded. To replace these participants, a second round of data collection was carried out using 55 new participants. These data were added to the overall sample and analyzed as described above. No additional participants were excluded following this analysis. The mean similarity ratings for all letter pairs can be found in the visual-similarity matrix in the [online supplementary materials](#), while Table 1 summarizes the numbers of pairs falling into each of 12 similarity bands. Most pairs were rated as fairly low in similarity, with the medians being 1.67 for uppercase and 1.70 for lowercase.

The main objective of constructing this matrix was to create a measure of *visual* similarity while ignoring *phonological* similarity. We were able to test whether participants' knowledge of the letter sounds exerted a strong influence on their responses by examining the ratings assigned to the two lowercase letters "a" (adult form) and "c" (infant form) and to the two uppercase letters "I" (serif) and "l" (sans serif). The five forms of the letter "a" that included diacritic marks were all shown using the infant form ("à", "á", "â", "ã", and "ö"). If participants were mainly influenced by visual similarity, the ratings given when these five letters were assessed with the infant form "c" should be higher than

**Table 1** Summary of letter visual-similarity ratings, classified by pair type and similarity band

Band <sup>a</sup>	Uppercase			Lowercase			Combined		
	N <sup>b</sup>	Pct <sup>c</sup>	SD <sup>d</sup>	N <sup>b</sup>	Pct <sup>c</sup>	SD <sup>d</sup>	N <sup>b</sup>	Pct <sup>c</sup>	SD <sup>d</sup>
1.00–1.50	542	40.9	0.13	583	42.3	0.13	1,125	41.6	0.13
1.51–2.00	340	25.6	0.15	243	17.6	0.14	583	21.6	0.15
2.01–2.50	157	12.0	0.15	124	9.0	0.14	281	10.4	0.14
2.51–3.00	72	5.3	0.15	102	7.4	0.14	174	6.4	0.14
3.01–3.50	42	3.2	0.15	94	6.9	0.14	136	5.0	0.15
3.51–4.00	31	2.3	0.17	84	6.1	0.16	115	4.3	0.16
4.01–4.50	44	3.3	0.12	48	3.4	0.13	92	3.4	0.13
4.51–5.00	23	1.7	0.12	22	1.6	0.16	45	1.7	0.14
5.01–5.50	8	0.6	0.12	13	1.0	0.12	21	0.8	0.12
5.51–6.00	5	0.4	0.12	10	0.7	0.17	15	0.6	0.16
6.01–6.50	37	2.8	0.12	42	3.1	0.13	79	2.9	0.13
6.51–7.00	25	1.9	0.06	13	0.9	0.07	38	1.4	0.06
Total	1,326	100.0		1,378	100.0		2,704	100.0	

<sup>a</sup> The rating scale is from 1 (*not at all similar*) to 7 (*very similar*), and this range has been divided into 12 bands. <sup>b</sup> The number of pairs that fall within each band. <sup>c</sup> The percentage of all of the pairs that fall within each band. <sup>d</sup> The standard deviation of the rating scores within each band

the ratings given when these five letters were assessed with the adult form “a”. In contrast, if participants were overly influenced by their knowledge of the phoneme associated with these letters, similar ratings should have been given when these five letters were assessed with the letters “a” and “o”. The same argument can be made for the ratings given to the letters “I” and “l”, as the three forms with diacritics were all shown in a sans serif font (“í”, “î”, and “ï”).<sup>3</sup>

Looking at the relevant data from the visual-similarity matrix (see the online supplementary materials), it can be seen that for the five letters “à”, “á”, “â”, “ã”, and “ä” higher similarity ratings were given when they were assessed with the infant form “a” (6.63, 6.53, 6.10, 6.00, and 6.35, respectively) as compared to when they were assessed with the adult form “a” (4.69, 5.07, 5.00, 4.30, and 4.67, respectively). As these five matched pairs represent a relatively small data set, rather than attempting to directly assess whether the differences between them were significant, we instead analyzed the individual participant ratings ( $n = 300$ ) that made up these ten means. A Mann–Whitney test revealed that the ratings given when the five letters “à”, “á”, “â”, “ã”, and “ä” were assessed with the letter “a” ( $M = 6.32$ ,  $Mdn = 6$ ) were significantly higher than

the ratings given when these letters were assessed with the letter “a” ( $M = 4.74$ ,  $Mdn = 5$ ),  $U = 5,733$ ,  $z = -7.68$ ,  $p < .001$ ,  $r = -.44$ . A similar pattern was observed for the ratings given to the three letters “í”, “î”, and “ï”, which were all shown in sans serif form. Higher ratings were given when these three letters were assessed with the sans serif “l” (6.50, 6.47, and 6.45, respectively) than when they were assessed with the serif “l” (6.27, 6.07, and 6.10, respectively). Analyzing the raw scores that made up these means ( $n = 180$ ) revealed that ratings given to the three letters “í”, “î”, and “ï” when assessed with the letter “l” ( $M = 6.47$ ,  $Mdn = 7$ ) were significantly higher than when they were assessed with the letter “l” ( $M = 6.14$ ,  $Mdn = 7$ ),  $U = 3,260$ ,  $z = -2.37$ ,  $p = .018$ ,  $r = -.18$ .

To summarize, while the analyses performed for lowercase “a” and uppercase “l” cannot rule out the possibility that participants were influenced by their knowledge of the letter sounds, we would not have expected to find the above-mentioned differences in the ratings if letter sound knowledge had dominated the participants’ assessments. Even if letter sound knowledge was partially influencing the ratings assigned, it was not strong enough to mask the effect of visual similarity, at least in the specific examples that we examined. These two results therefore suggest that participants were mostly ignoring their knowledge of the letter sounds and were assigning ratings based on the visual similarity of the stimuli, as was required of them. These results are consistent with the findings from Treiman, Levin, and Kessler (2007), who constructed a letter visual-similarity matrix for Hebrew letters using the same procedure reported here—that of adult participants providing similarity ratings to written letter pairs. Treiman, Levin, and Kessler (2007) also attempted to verify that participants had ignored their knowledge of letter sounds when rating the visual similarity

<sup>3</sup> Of the five forms of the letter “a” with diacritic marks and the three forms of the letter “l” with diacritic marks that were included in the experiment, only the forms with the grave accent, “á” and “î”, exist in Spanish. The grave accent in Spanish can be used with all five vowel letters, and it is used to mark stress in words not following the default stress pattern. However, the grave accent does not change the phoneme associated with the letter. Therefore, had the participants, virtually all native Spanish speakers, taken into account the phonemes associated with the letters “a” and “l”, they would likely have associated the same phonemes with all occurrences of “a” and “l”, both with and without diacritic marks.

of the Hebrew letters, although this was done using a method different from the one that was used in the present study. In particular, Treiman, Levin, and Kessler (2007) asked a second set of participants who were non-Hebrew speakers, and so were unfamiliar with the actual sounds associated with each symbol, to rate the same set of letters. The correlation between the two groups of participants was  $r = .86$ . Treiman, Levin, and Kessler (2007) concluded that “this correlation gives us reason to believe that the ratings of the Israeli participants reflect, for the most part, characteristics of letters’ visual forms that are salient regardless of a viewer’s familiarity with the letters” (p. 97). Our analyses of lowercase “a” and uppercase “I” are in agreement with the conclusions derived by Treiman, Levin, and Kessler (2007), ensuring that the participants in our study largely ignored letter sound knowledge when assessing the letter pairs.

As was already mentioned, the main aim of this study was to create an instrument that would serve investigators carrying out reading research in different alphabetic scripts. For this reason, we included a large number of stimuli in an effort to gain as wide a language coverage as possible. This meant that 44 % of the stimuli used in the experiment contained a diacritic mark. Examining the pairs from Table 1 that make up bands 6.01–6.50 and 6.51–7.00 (i.e., the most visually similar pairs) revealed that only three of 117 pairs did not contain a diacritic mark. Given this, might the inclusion of the diacritic marks somehow bias participants’ ratings and make them unsuitable for use in languages that do not include them? For example, in English, the letters “E” and “F” are visually similar. However, our study also included the letters “È”, “É”, “Ê”, and “Ë”. Might the presence of these four letters have biased participants such that they rated the combination “E F” as being less similar than they otherwise would have if the four forms of the letter “E” with diacritic marks had not been present? To test this possibility, we ran an additional study. In this second study, 60 participants were required to rate the visual similarity of either 131 lowercase pairs or 100 uppercase letter pairs (30 participants for each case). The participants in this second study were all native Spanish speakers who were also undergraduates from the University of Granada. The presentation and procedure were identical to those of the first study. However, there was one major difference: The letter pairs were formed using only the 26 letters of the English alphabet (all of which exist in the Spanish alphabet), with no diacritic marks included. By comparing the ratings from the two studies, we could gauge whether the inclusion of diacritic marks in the first study biased the ratings given by the participants in the first study. Calculating the Spearman’s rank correlation coefficient revealed high correlations between the two studies for both the uppercase ( $r_s = .91, p < .001$ ) and lowercase ( $r_s = .92, p < .001$ ) letters. Additionally, Wilcoxon signed-rank tests revealed no significant differences

between the ratings in Studies 1 and 2 for either uppercase ( $M_{S1} = 2.52, Mdn_{S1} = 2.15; M_{S2} = 2.50, Mdn_{S2} = 2.08$ ),  $z = -.539, p = .590, r = -.04$ , or lowercase ( $M_{S1} = 2.49, Mdn_{S1} = 1.93; M_{S2} = 2.49, Mdn_{S2} = 1.93$ ),  $z = -.479, p = .632, r = -.03$ . These results indicate that including diacritic marks in the first study did not bias the way in which participants rated letter combinations without diacritic marks.

Another possibility is that cultural and/or linguistic differences could influence the ratings assigned by participants. To test this possibility, we ran a third study. Sixty participants, all native English-speaking undergraduates from the School of Psychology in Bangor, Wales, were asked to rate the visual similarity of either 125 lowercase or 132 uppercase letter pairs (30 participants for each case). Once again, the procedure and instructions were the same, and the letter pairs were again constructed using only the 26 letters of the English alphabet, with no diacritic marks included. By comparing the ratings from Study 1 with those from Study 3, we could determine whether English and Spanish speakers assessed the visual similarity of letter pairs in the same way. Spearman’s rank correlation coefficient revealed high correlations between the two studies for both the uppercase ( $r_s = .95, p < .001$ ) and lowercase ( $r_s = .93, p < .001$ ) letters. Additionally, Wilcoxon signed-rank tests revealed no significant difference between the ratings given in Studies 1 and 3 for lowercase letters ( $M_{S1} = 2.35, Mdn_{S1} = 1.83; M_{S3} = 2.38, Mdn_{S3} = 1.77$ ),  $z = -.430, p = .667, r = -.03$ . However, for the uppercase letters, the English ratings were significantly higher than the Spanish ratings ( $M_{S1} = 2.31, Mdn_{S1} = 1.94; M_{S3} = 2.47, Mdn_{S3} = 2.13$ ),  $z = -5.526, p < .001, r = -.34$ ). While the latter result may be an indication that cultural and/or linguistic differences between the participants influenced the ratings assigned, we do not believe that this is the case, for two reasons. Firstly, if such a difference existed, we would expect to see systematic differences between the two groups of participants for both upper- and lowercase letters, yet there was no significant difference in the lowercase ratings. Secondly, when we looked closely at the individual data for uppercase letters, between 15 % and 30 % of the ratings of three of the English participants were 2 *SDs* above the mean ratings. When these participants were removed from the analyses, the difference between the two groups was no longer significant, ( $M_{S3} = 2.37, Mdn_{S3} = 1.98$ ),  $z = -1.32, p = .188, r = -.08$ . It is worth noting that each participant in Study 3 rated all possible pairings, whereas in Study 1 each participant rated fewer than 10 % of all possible letter pairings. As a result, atypical participants in Study 3 would have had a much larger impact on the ratings than would atypical participants in Study 1. Thus, the significant difference between the two groups found for just uppercase letters seems to have been caused by outlier ratings within the English group. Taken together, we argue that the results of comparing the three studies validate our

suggestion that the letter visual-similarity matrix that we introduced will be applicable for reading research carried out in a large number of languages, despite the fact that ratings in the present study were provided almost exclusively by native Spanish speakers.

To investigate how our matrix differs from those constructed using different paradigms, we compared our matrix to matrices based on (a) direct ratings of letter pair similarities (i.e., a similar paradigm as the one used in the present study; Boles & Clifford, 1989); (b) feature analysis of letter forms (i.e., nonbehavioral data; Briggs & Hocevar, 1975); (c) low contrast/noisy presentation of letters (Liu & Arditi, 2001); (d) brief presentation of letters (Mueller & Weidemann, 2012); and (e) presentation of small letters (Phillips et al., 1983). The results are summarized in Table 2.

It is worth noting two points about the correlations presented in Table 2. Firstly, while the matrix presented in the present study correlated significantly with all six of the matrices that we examined, the highest correlation was found between our matrix and the matrix that was constructed using a similar paradigm (Boles & Clifford, 1989). Secondly, correlations between studies using different paradigms tended to be weaker, and occasionally nonsignificant. This is an important point, and it suggests that the various methods used to increase the numbers of errors made by participants do not produce consistent results. This is not a problem in itself, as each technique was deployed to satisfy specific research questions. However, this observation supports the argument that matrices constructed using atypical reading data obtained from speeded reading-aloud tasks and/or under degraded presentation conditions may be inadequate for use by researchers who aim to determine the factors that underlie our capacity to read in daily life.

The matrix presented in the present article may prove very useful to reading researchers, yet one of its limitations is that the ratings were gathered using printed letters. Fonts displayed on computer screens can differ from their printed equivalents, due to the technical issues involved in rendering images on a computer screen, such as subpixel shading and anti-aliasing. For this reason, the ratings contained in the present matrix may be unsuitable for research carried out using computer-based rather than printed stimuli. Additionally, the ratings may be unsuitable for research in which stimuli are presented in fonts that are markedly different from the ones used here. A further limitation of the present matrix is that it lacks similarity ratings between lowercase and uppercase letters (e.g., comparing “B” to “b”). Therefore, it may not be useful to reading researchers who seek to control for such comparisons in the experimental paradigms that they are using (e.g., Kinoshita & Kaplan, 2008).

## Conclusion

Although many letter visual-similarity matrices have been published over the last century (for a review, see Mueller & Weidemann, 2012), these can only be used to carry out relevant research in English. Furthermore, these matrices have mostly been formed from data generated in atypical reading conditions, using, for example, speeded naming or degraded presentation conditions. For these reasons, the letter visual-similarity matrices published to date are of limited utility for research undertaken in alphabetical languages other than English, and are also unsuitable for reading researchers who are typically interested in the mental processes occurring in natural reading conditions. The aim of the present article was to

**Table 2** Summary of Spearman’s rank correlation coefficients between the letter visual-similarity matrices from different studies

	2	3	4 <sup>a</sup>	5 <sup>b</sup>	6 <sup>b</sup>	7 <sup>a</sup>
1. Present Study	.735 <sup>***c</sup>	.668 <sup>***d</sup>	.328 <sup>***d</sup>	-.201 <sup>***d</sup>	-.173 <sup>**d</sup>	.310 <sup>***d</sup>
2. Boles & Clifford (1989) (direct rating of letter pairs)	–	.575 <sup>***e</sup>	.359 <sup>***d</sup>	-.262 <sup>***d</sup>	-.059 <sup>d</sup>	.286 <sup>***e</sup>
3. Briggs & Hocevar (1975) (letter feature analysis)		–	.330 <sup>***e</sup>	-.208 <sup>***e</sup>	-.180 <sup>**e</sup>	.338 <sup>***d</sup>
4. Liu & Arditi (2001) <sup>a</sup> (low contrast presentation)			–	-.113 <sup>*d</sup>	-.123 <sup>*d</sup>	.355 <sup>***e</sup>
5. Mueller & Weidemann (2012) Exp. 1 <sup>b</sup> (brief presentation)				–	.129 <sup>*d</sup>	-.141 <sup>*e</sup>
6. Mueller & Weidemann (2012) Exp. 2 <sup>b</sup> (brief presentation)					–	-.030 <sup>e</sup>
7. Phillips et al. (1983) <sup>a</sup> (reducing letter size)						–

<sup>a</sup> The data from these studies consist of two scores for each letter pair (Letter 1 [L1] as the target and Letter 2 [L2] as the incorrect response, and vice versa). We calculated an average of these two scores to use in the present correlations. <sup>b</sup> Correlations for the Mueller and Weidemann (2012) study were based on the *predicted accuracy scores* of their best model, which is described by the parameters contained in their Table 5 (Mueller & Weidemann, 2012, p. 31). The scores represent the predicted probability of a participant accurately identifying a target letter (L1) in the presence of a foil letter (L2). The averages of these two probabilities across the two orders of the letters were calculated and used in the present correlations. As these values represent predicted accuracy scores, higher values would suggest that the target and foil were not visually similar, while lower values would suggest that the target and foil were visually similar. For this reason, we would expect a negative correlation between this study and the other visual-similarity matrices contained in the table. <sup>c</sup> Upper- and lowercase letter pairings,  $n = 650$ . <sup>d</sup> Uppercase letter pairings only,  $n = 325$ . <sup>e</sup> Uppercase letter pairings only, excluding the letter “T” due to font differences,  $n = 300$ . \*  $p < .05$ . \*\*  $p < .005$ . \*\*\*  $p < .001$



overcome these issues by presenting a letter visual-similarity matrix that is based on similarity ratings collected from untimed responses to clearly presented upper- and lowercase letters. Additionally, the letter similarity ratings included in our matrix have been obtained using letters that are present in a substantial number of Latin-based languages, including Catalan, Dutch, English, French, Galician, German, Italian, Portuguese, and Spanish. To our knowledge, this is the first time that letter visual-similarity data have become available for such a wide range of alphabets. Furthermore, the items included in the present matrix are not limited to just the standard letters of these languages, but also include letters with diacritic marks. Although these forms may not be considered separate letters, they nevertheless affect stress assignment, pronunciation, and the meanings of words in the languages in which they occur, highlighting the importance of their inclusion in a letter visual-similarity matrix.

We believe that our matrix will be a valuable tool for conducting experimental research in psycholinguistics, which may involve the selection of letter stimuli on the basis of their visual similarity (Burgund & Abernathy, 2008); for informing cognitive neuropsychological reading research (Brunsdon, Coltheart, & Nickels, 2006; Fiset et al., 2006); and for determining the factors that influence reading acquisition and development (Treiman et al., 2006, 2012; Treiman, Levin, and Kessler, 2007).

**Author note** P.M. is now at Royal Holloway, University of London. The research leading to these results received funding from the [European Community's] Seventh Framework Programme (FP7/2007-2013) under Grant No. ELDEL PITN-GA-2008-215961-2. Funding was also provided by Project PSI2010-21983-C02-01, Ministerio de Ciencia e Innovación, Gobierno de España, and the HUM-820 group of the Junta de Andalucía.

## References

- Bernard, M. L., Chaparro, B. S., Mills, M. M., & Halcomb, C. G. (2002). Examining children's reading performance and preference for different computer-displayed text. *Behaviour and Information Technology*, 21, 87–96. doi:10.1080/01449290210146737
- Bernard, M. L., Chaparro, B. S., Mills, M. M., & Halcomb, C. G. (2003). Comparing the effects of text size and format on the readability of computer-displayed Times New Roman and Arial text. *International Journal of Human Computer Studies*, 59, 823–835. doi:10.1016/S1071-5819(03)00121-6
- Boles, D. B., & Clifford, J. E. (1989). An upper- and lowercase alphabetic similarity matrix, with derived generation similarity values. *Behavior Research Methods*, 21, 579–586. doi:10.3758/BF03210580
- Briggs, R., & Hocevar, D. J. (1975). A new distinctive feature theory for upper case letters. *The Journal of General Psychology*, 93, 87–93.
- Brunsdon, R., Coltheart, M., & Nickels, L. (2006). Severe developmental letter-processing impairment: A treatment case study. *Cognitive Neuropsychology*, 23, 795–821. doi:10.1080/02643290500310863
- Burgund, E. D., & Abernathy, A. E. (2008). Letter-specific processing in children and adults matched for reading level. *Acta Psychologica*, 129, 66–71. doi:10.1016/j.actpsy.2008.04.007
- Fiset, D., Blais, C., Arguin, M., Tadros, K., Éthier-Majcher, C., Bub, D., & Gosselin, F. (2009). The spatio-temporal dynamics of visual letter recognition. *Cognitive Neuropsychology*, 26, 23–35. doi:10.1080/02643290802421160
- Fiset, D., Gosselin, F., Blais, C., & Arguin, M. (2006). Inducing letter-by-letter dyslexia in normal readers. *Journal of Cognitive Neuroscience*, 18, 1466–1476. doi:10.1162/jocn.2006.18.9.1466
- 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
- Kinoshita, S., & Kaplan, L. (2008). Priming of abstract letter identities in the letter match task. *Quarterly Journal of Experimental Psychology*, 61, 1873–1885. doi:10.1080/17470210701781114
- Kuennapas, T., & Janson, A.-J. (1969). Multidimensional similarity of letters. *Perceptual and Motor Skills*, 28, 3–12. doi:10.2466/pms.1969.28.1.3
- Liu, L., & Arditi, A. (2001). How crowding affects letter confusion. *Optometry and Vision Science*, 78, 50–55.
- Mueller, S. T., & Weidemann, C. T. (2012). Alphabetic letter identification: Effects of perceivability, similarity, and bias. *Acta Psychologica*, 139, 19–37. doi:10.1016/j.actpsy.2011.09.014
- Phillips, J., Johnson, K., & Browne, H. (1983). A comparison of visual and two modes of tactual letter resolution. *Attention, Perception, & Psychophysics*, 34, 243–249. doi:10.3758/BF03202952
- Reich, L. N., & Bedell, H. E. (2000). Relative legibility and confusions of letter acuity targets in the peripheral and central retina. *Optometry and Vision Science*, 77, 270–275.
- Snow, C. E., Burns, M. S., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academies Press.
- Taslim, J., Wan Adnan, W., & Abu Bakar, N. (2009). Investigating children preferences of a user interface design. In J. A. Jacko (Ed.), *Human-computer interaction: New trends. Proceedings of the 13th International HCI Conference, Part I (Vol. 5610, pp. 510–513)*. Berlin: Springer.
- Treiman, R., Cohen, J., Mulqueeny, K., Kessler, B., & Schechtman, S. (2007). Young children's knowledge about printed names. *Child Development*, 78, 1458–1471. doi:10.1111/j.1467-8624.2007.01077.x
- Treiman, R. A., & Kessler, B. (2004). The role of letter names in the acquisition of literacy. In R. Kail (Ed.), *Advances in child development and behavior* (Vol. 31, pp. 105–135). Orlando: Academic Press.
- Treiman, R., Kessler, B., & Pollo, T. C. (2006). Learning about the letter name subset of the vocabulary: Evidence from US and Brazilian preschoolers. *Applied Psycholinguistics*, 27, 211–227. doi:10.1017/S0142716406060255
- Treiman, R., Levin, I., & Kessler, B. (2007). Learning of letter names follows similar principles across languages: Evidence from Hebrew. *Journal of Experimental Child Psychology*, 96, 87–106. doi:10.1016/j.jecp.2006.08.002
- Treiman, R., Levin, I., & Kessler, B. (2012). Linking the shapes of alphabet letters to their sounds: The case of Hebrew. *Reading and Writing*, 25, 569–585. doi:10.1007/s11145-010-9286-3
- Walker, S., & Reynolds, L. (2003). Serifs, sans serifs and infant characters in children's reading books. *Information Design Journal*, 11, 106–122. doi:10.1075/idj.11.2.04wal
- Walker, S., & Thiessen, M. (2011). Typography for children's reading. *Fonts in Focus*, 8, 10–11.
- Wilkins, A., Cleave, R., Grayson, N., & Wilson, L. (2009). Typography for children may be inappropriately designed. *Journal of Research in Reading*, 32, 402–412. doi:10.1111/j.1467-9817.2009.01402.x