The acceptability, perception, and production of novel Chinese character components

James Myers
https://orcid.org/0000-0002-3866-1969
Graduate Institute of Linguistics
National Chung Cheng University
168 University Road, Min-Hsiung
Chia-Yi 621301
TAIWAN
Email: Lngmyers@ccu.edu.tw
Web: https://lngmyers.ccu.edu.tw/

Keywords: Writing systems; Chinese characters; acceptability; lexical neighbors; visual perception; handwriting

Abstract word count: 233 words

Word count (not including references): 10,275

Authors' contributions: James Myers: Conceptualization, Methodology, Software, Formal Analyses, Data curation, Writing – original draft preparation, Writing – reviewing and editing, Funding acquisition.

Data accessibility:
https://osf.io/x6btj/overview?view_only=f3709e9293b945beabf74f300c24b52e

Funding: This study was supported by the National Science and Technology Council (NSTC) of Taiwan under grants 109-2410-H-194-096-MY3 and 112-2410-H-194-030-MY3.

Conflicts of interest: None.

Ethics approval and consent: This study was conducted as part of projects approved by the National Chung Cheng University human research ethics center, which confirmed that the study conforms to the Declaration of Helsinki and relevant laws of Taiwan, including those relating to informed consent and Taiwan's Personal Data Protection Act.

Material from other sources: None.

Acknowledgements: This study was made possible by funding from the NSTC, five lab assistants (Chen Yangru, Liu Meichun, Chang Yuchu, Hsu Shaofei, Lin Wenchi), and numerous experimental participants. The stimuli used BabelStone fonts created and maintained by Andrew West: https://www.babelstone.co.uk/Fonts/. An early version of this study was presented at the 2024 meeting of the Linguistic Society of New Zealand.

See what the reviewers are saying!

* “This is a nice manuscript but in my opinion it is not a psycholinguistic paper.”

* “...this particular writeup seems to be designed mainly for a readership that is familiar with these debates and designs in writing research and perhaps one in applied psycholinguistics.”

* “Why are only single-component character used in the research materials ?”

* “Strictly speaking, all the stimuli used were non-Chinese materials, and all the tasks used were not natural reading and writing tasks.”

The acceptability, perception, and production of novel Chinese character components

Abstract

Virtually all research on productive orthographic knowledge has focused on glyph combinations, like letter sequences, but glyph inventories show formal patterns as well. To explore reader/writer knowledge of such patterns, Chinese-speaking participants were given a variety of tasks using a set of non-lexical character components (radicals or logographemes). In Experiment 1, binary acceptability judgments were collected for these items and then modeled in terms of pretest-derived lexical neighbor scores, plus variables associated with visual perception and manual articulation. Neighbor scores had a significant positive influence on acceptability, even with the sensorimotor variables factored out. In Experiment 2, participants performed a same-different task, with the latencies for correct responses to same pairs modeled using the same predictors as Experiment 1, plus that experiment's by-item acceptance rates. While acceptability had a marginally significant facilitative effect, neighbor scores had no effect, despite the lack of collinearity. In Experiment 3, participants performed a handwritten copying task. Acceptability had significant facilitative effects on accuracy and response latency, again without an additional effect of neighbor scores. Together these findings demonstrate the productivity of glyph form knowledge in Chinese, and also suggest that acceptability taps into the same processes involved in perception and production without itself being reducible to simple perceptual, production, or lexical variables. The results highlight the richness of glyph form knowledge and the processes underlying acceptability judgments, both of which remain woefully understudied in all writing systems.

Keywords: Writing systems; Chinese characters; acceptability; lexical neighbors; visual perception; handwriting

Word count (not including references): 10,275

1. Introduction

Literate readers/writers do not just memorize written words by rote, but also have expectations about combinations of glyphs (here this term is used to refer to basic graphic forms, like letters). This productive knowledge, mostly implicit, has been well established for alphabetic systems (Chetail, 2015; Treiman & Kessler, 2022) and for component combinations within Chinese characters (Tong & McBride-Chang, 2010; Wang, Yang, Shu, & Zevin, 2011). However, the inventories of letters or Chinese character components themselves also display internal formal regularities (Myers, 2019; Primus, 2004; Wang, 1983; Watt, 1988), raising the possibility that knowledge of these regularities is also productive. This study explores this possibility by analyzing acceptability judgments for novel (non-lexical) Chinese character components in terms of lexical similarity, perception, and production variables, and then examining the roles that all of these variables play in real-time perception and production. While the study was conducted on Chinese script, the theoretical challenges that it raises are relevant to all writing systems, where glyph form processing remains woefully understudied.

1.1 Chinese character structure

There are three levels of Chinese character structure. At the top are the characters themselves, which as the system's graphemes (Meletis, 2019) generally represent monosyllabic morphemes, like 媽 *mā* 'mother' (traditional characters are cited because this study was conducted in Taiwan). At the intermediate level are the system's basic glyphs, namely the character-internal components (including potentially interpretable radicals and purely formal logographemes: Chen & Cherng 2013; Law & Leung 2000); for example, 媽 contains the semantic radical 女 (*nǚ* 'female') and the phonetic component 馬 (*mǎ* 'horse'). Components need not have synchronic meanings or pronunciations, let alone form full characters; for example, 色 (*sè* 'color') and 角 (*jiǎo* 'horn') share the synchronically meaningless and unpronounceable bound logographeme ⺈, combined respectively with 巴 (*bā* 'python [classical]') and 用 (*yòng* 'use'). Finally, at the lowest level of character structure are the strokes; for example, the component 女 contains the basic stroke 一 and the complex (multi-axis) stroke 𡿨.

Virtually all psycholinguistic and neurolinguistic studies on Chinese characters have focused on the relationship between the character and component levels (see reviews in Duan & Cai, 2024; Wu, 2022; Zhang & Ke, 2018). This research has established that components play important roles in character learning (Tong, Tong, & McBride, 2017), recognition

(Hsieh, Lin, Li, & Wu, 2021), and handwriting (Chen & Cherng, 2013; Lau, 2020; cf. Wang, Huang, Zhou, & Cai, 2020). Moreover, readers/writers also demonstrate productive knowledge of what makes for a possible character, based on how they treat two types of non-lexical characters: pseudocharacters (accidental gaps with lexical components in their allowed positions and allographic forms) and non-characters (systematic gaps, including characters with lexical components in the wrong positions or allographic forms). Literate adults (Wang et al., 2011; Wu, 2022) judge pseudocharacters as more acceptable than non-characters, and process them faster, more accurately, and with less neural effort; children gradually develop these skills as well (Shu & Anderson, 1999; Tong & McBride-Chang, 2010). Because the processing of predictable properties need not depend solely on rote memory, expectations about the systematic nature of character structure may be what makes it possible to learn, recognize, and handwrite the thousands of characters mastered by the typical adult reader/writer (Hue, 2003).

Given this, it is curious how little attention has been paid to the fact that Chinese character components themselves number in the hundreds (Kordek, 2013). Could it be that readers/writers also exploit their expectations about systematic regularities in component form? Such regularities do in fact abound. Basic strokes always run along one of the canonical axes (horizontal or vertical) or the two main diagonals, as in 米 (*mǐ* ‘uncooked rice’), and can be modified in only two ways, via curving or hooking, as in the two long strokes at the bottom of 亦 (*yì* ‘also’). Complex strokes are always formed by linking basic ones, as on the right of 乃 (*nǎi* ‘be’, with the complex stroke ㇋ stringing together the basic strokes 一丨一亅). There are also conventions concerning how strokes are ordered when handwriting components, generally from left to right and top to bottom (Wan, 2017), as well as a variety of subtler but robust regularities linking stroke form and position (Myers, 2019; Wang, 1983). For example, the characters 亦 and 用 conform to the generalization that the curved vertical stroke can only appear on the left, while the left-hooked stroke never can; complementarily, right-hooked vertical strokes are restricted to the left, as in 衣 (*yī* ‘clothing’). The characters 工 (*gōng* ‘work’) and 土 (*tǔ* ‘soil’) are regular in another way, since in both the lower horizontal stroke is the longer one; the latter character contrasts with 士 (*shì* ‘scholar’), but such violations of the generalization are rare (Yang & Wang, 2018).

Glyph-level form regularities are hardly unique to Chinese; every script has its own unique “look.” Within any given inventory, glyphs can be distinguished by means of a relatively small number of binary visual features, albeit more than those needed to discriminate among phonemes (Kim, Allassonnière-Tang, Miton, & Morin, 2025). Even though Roman letters would seem to be few enough in number to memorize solely by rote, they are decomposable into an even smaller small set of basic stroke types (Primus, 2004; Watt, 1988), which obey simple principles of combination and position; for example, uppercase Roman letters are more likely to “face” rightward (<B>) than leftward (<J>), and all letters are more likely to contain vertical strokes (<B, b>) than horizontal ones (<E, e>). It may be, then, that the human mind expects even such small glyph systems to obey formal regularities in order for them to be efficiently learned and processed.

1.2 Productive knowledge of glyph form patterns

It is as yet virtually unknown whether such formal glyph-internal patterns are mentally active, since research on productive orthographic knowledge has focused almost exclusively on non-lexical combinations of lexical glyphs, like non-lexical letter strings in studies on English speakers (Chetail, 2015; Treiman & Kessler, 2022) and studies comparing pseudocharacters and non-characters in Chinese (Wu, 2022). A nontrivial number of studies have included non-lexical glyphs, but their focus still remains on the glyph combinations, not the glyphs themselves (He & Tong, 2017; Rastle, Lally, Davis, & Taylor, 2021; Taylor, Plunkett, & Nation, 2011; Vidal, Viviani, Zoccolan, & Crepaldi, 2021), or at most the mappings between glyph features and phonological features (Byrne & Carroll, 1989) rather than the form patterns in their own right. Even when two-dimensional visual images rather than linear strings have been tested, the patterns still involve the layout of discrete elements (Fiser & Aslin, 2001; Westphal-Fitch, Giustolisi, Cecchetto, Martin, & Fitch, 2018; Vidal et al., 2021) rather than sorts of “natural” glyph patterns reviewed above, which typically involve stroke features or their interactions with position.

Nevertheless, there have been a smattering of studies on glyph forms as forms. Most have taken a broad view, discussing glyph form evolution in terms of general principles like visual complexity, symmetry, or stroke axis (Changizi & Shimojo, 2005; Han, Kelly, Winters, & Kemp, 2022; Morin, 2018; Miton & Morin, 2021), but some have looked at the productive knowledge of glyph form regularities within individual scripts, as is done in the present study. While simple Roman letter form characteristics like symmetry or the number of strokes explain surprisingly little of the variance in young English-speaking children’s handwriting accuracy (Puranik, Petscher, & Lonigan, 2014), children are nevertheless sensitive to formal systematicities within the letter inventory, most notably tending to mirror-reverse left-facing letters in order to conform to the right-facing generalization (Treiman, Gordon, Boada, Peterson, & Pennington, 2014). Adults also retain their productive knowledge of Roman letter form. Jameson (1994) found that English-speaking readers rated non-letters as more acceptable the better they conformed with the generalizations of Watt (1988), and Reinken (2023) demonstrates that handwritten German letters show formal regularities consistent with the letter-internal structure posited by Primus (2004), including giving primacy to vertical strokes (e.g., in fluent handwriting, <b, f, h, r, t> may all be reduced to <l>).

A few similar observations have been made for Chinese script. In a series of acceptability judgment experiments, Myers (2019) found higher acceptability for non-lexical stroke combinations that conformed to the curving, hooking, and lengthening patterns reviewed above, while Xin and Lyu (2022) used non-lexical stroke combinations to test the productivity of conventional stroke order principles. However, both studies were hindered by the relatively unnaturalness of their stimuli (ad hoc combinations of strokes), and were only able to test a few of the more obvious generalizations, likely missing much of what fluent readers/writers know about component form. It is thus necessary not only to extend this work further, but to do so more carefully, in order to form a solid basis that can be built on in the study of glyph form processing more generally.

1.3 Testing productive form knowledge

For spoken and signed phonology, the most straightforward way to test productive knowledge is simply to ask participants to judge whether, or to what degree, non-lexical test items (like *blick* or *bnick* for English speakers; Chomsky & Halle, 1965) seem acceptable, or wordlike, or well-formed (Arendsen, van Doorn, & de Riddera, 2010; Bailey & Hahn, 2001; Cohn, Fougeron, & Huffman, 2017; Frisch & Zawaydeh, 2001; Gouskova & Becker, 2013; Hayes & White, 2013; Kawahara, 2011). Phonological acceptability judgments are strongly affected by lexical variables like neighborhood density, which reflects the holistic similarity of a non-lexical test item with real words, and phonotactic probability, which represents the degree to which a non-lexical test item reflects morpheme-internal regularities (see Bailey & Hahn, 2001, for a review); phonotactic probability is in turn shaped, at least diachronically, by sensorimotor processing (Dziubalska-Kołaczyk, 2014). The same lexical and sensorimotor variables also affect real-time processing: non-lexical items with lower neighborhood density and/or phonotactic probability are processed faster and more accurately in the lexical decision task (Gong, Zhang, & Fiorentino, 2024), the same-different judgment task (Vitevitch & Luce, 1999), and the repetition naming task (Vitevitch & Luce, 2005). The independent influence of typological frequency and/or phonetic motivation on acceptability judgments can be distinguished from the effects of lexical variables by controlling them separately (Hayes & White, 2013). There is some evidence that such universal factors are not merely sensorimotor but more abstractly linguistic, even for perception tasks (Berent & Lennertz, 2010).

Because the present study includes the acceptability judgment task as one of its methods, it is important to say a bit more about its nature. The making of phonological acceptability judgments does not involve real-time consultation of a centralized store of phonological knowledge, just as mental grammar more generally is not an isolated module but rather a higher-level description of processing itself (Lewis & Phillips, 2015; Neeleman & van de Koot, 2010). Acceptability judgments merely reflect the relative ease of this processing as a whole (Gross, 2020; Topolinski & Strack, 2009). These ideas are consistent with neuroimaging evidence that the making of phonotactic acceptability judgments is accompanied by neural activation distributed widely across sensorimotor areas and networks associated with lexical access (Avcu, Newman, Ahlfors, & Gow Jr, 2023). The broad overview of the word form processing system provided by acceptability judgments may explain why they cannot be completely reduced to well-established variables like neighborhood density or phonotactic probability (Bailey & Hahn, 2001; Frisch & Zawaydeh, 2001; Gouskova & Becker, 2013); much of the variance in acceptability remains unexplained, and unfortunately, still largely unexplored.

Just as with spoken words, the processing of glyph form is expected to be influenced by both universal and lexical factors. Regarding the former, Morin (2018) demonstrates that unrelated writing systems share a number of properties seemingly shaped by visual biases, including the favoring of strokes along the two cardinal axes (horizontal and vertical) over diagonals, the disfavoring of mixtures of the two types, and the preference for reflective symmetry along the vertical rather than the horizontal axis. Changizi and Shimojo (2005) suggest that such visual biases have had a greater influence on the historical development of orthographic form than motoric biases, because, among other reasons, any given text is only written once but may be read many times.

Investigations into the drawing of simple geometric figures demonstrates that there are also robust motoric constraints on handwriting. These include the preference for writing downward rather than upward and, for right-handers, writing rightward rather than leftward (Van Sommers, 1984). Writing along the horizontal axis is also easier to sustain than writing along the vertical axis, given that the former involves more robust muscles in the wrist while the latter involves the smaller and less resilient muscles in the fingers (Kushki, Schwellnus, Ilyas, & Chau, 2011). A subtler motoric bias is that when drawing a T-shaped stroke configuration, people prefer to anchor the starting point of the new stroke on the previously drawn stroke (e.g., in <T>: — before |). This bias, when combined with the downward and (right-handed) rightward biases, explains why studies consistently find that configurations like ┬ and ├ are easier to draw than configurations like ┤ and ┴ (Goodnow & Levine, 1973; Van Sommers, 1984). Still other universal factors are expected to affect both readers and writers, in particular the number and complexity of strokes, which challenge both visual and motoric processing.

A test item's similarity to glyphs in the lexical inventory should also affect how it is processed. However, it turns out that computing lexical glyph similarity is a nontrivial matter, quite unlike the well-established methods used in the study of glyph combinations. Because alphabetic writing systems reflect phoneme sequences in the spoken language, neighborhood density and phonotactic probability can be readily adapted to letter sequences, and indeed, such measures do affect how written words are accessed and produced (Chetail, 2015), how written nonwords are judged for acceptability (Bailey & Hahn, 2001), and how well words are learned (Storkel & Lee, 2011). For Chinese characters, similar observations have been made regarding the number of components shared with other characters (Li, Bi, Wei, & Chen, 2011). For individual glyphs, however, it is unclear how to define either lexical neighbors or orthotactic probabilities. Strokes within letters or within Chinese character components are not arranged linearly like letters, and while both the pedagogical and scientific literature tend to imply that Chinese stroke order is fixed, actual writers show considerable variation (Katayama, Uchida, & Sakoe, 2009; Wang, 1983; Yin, 2016). In any case, the two-dimensional arrangement of strokes within glyphs is intrinsically nonlinear; for example, letters pairs like <b, p> and <d, q> contain the same basic strokes written in the same order but combine them differently. Similarly, the Chinese characters 工 and 土 consist of the same strokes written in the same order, but contrast in the type of stroke contact (T-shaped vs. plus-shaped). In short, when studying individual glyph forms rather than glyph combinations, even the calculation of lexical similarity requires a new approach; as with so much in this research area, the prior literature provides very little guidance.

1.4 The present study

This study presents a series of experiments testing the acceptability, perception, and production of non-lexical Chinese character components, in order to explore what Chinese readers and writers expect about character component form, and to determine if this productive form knowledge goes beyond lexical similarity and universal sensorimotor biases.

Regression-based designs were used to avoid having to force largely unexplored variation into the few factors that happen to have been noticed by linguists (as was done in Myers, 2019). For similar reasons, Bailey and Hahn (2001) also adopted regression-based designs in their phonotactic acceptability judgment experiments (see also Baayen, 2010, and Keuleers & Balota, 2015, for advantages of this kind of design for lexical research more generally).

Just as phonotactic acceptability experiments use non-lexical stimuli that simulate monomorphemic words, all of the experiments here used the same set of novel single-component forms (i.e., similar to free components like 巴 rather than bound ones like ⺈). However, it is far harder to create non-lexical glyphs than non-lexical glyph combinations or phoneme sequences. Freely combining strokes (as in Myers, 2019, and Xin & Lyu, 2022) yields forms more geometrical than orthographic (cf. English listeners' perception of African clicks not just as sounding unlike English, but unlike speech: Best, 2020). It would also not do to create non-lexical forms by modifying one or two strokes in real character components, because this would make the sources obvious while leaving much of the combinatorial space unexplored. Creating stimuli through a stroke-combining grammar (analogous to the phoneme-combining grammar of Bailey & Hahn, 2001, or the algorithm for combining Chinese character components created by Chang, Tseng, Perfetti, & Chen, 2022) would just beg the question of what the psychologically real grammar might be for Chinese character glyphs.

To avoid all of these problems, this study exploited a naturally evolved set of non-Chinese glyphs used around 900-1200 CE to write the extinct Mongolic language Kitan (also known as Khitan or Liao), which was spoken in what is now Mongolia and parts of northern China and southeastern Russia (Kane, 2009). Kitan was written with two distinct scripts, a logographic "large" script and a quasi-syllabic "small" script (the latter arranged in squares). As fascinating as the Kitan scripts are, all that is relevant here is that both were strongly influenced by single-component Chinese characters, so much so that when printed in modern Chinese fonts, they appear Chinese-like to various degrees (see examples in Table 1). The test items, then, while not created specifically for these experiments, are orthographically plausible, having actually been used by a literate culture, and Sinoform without being actual Chinese. The full set of stimuli is given in the supplementary material (accessible via the OSF link provided before the reference list).

Due to the difficulties reviewed above in defining the similarity of non-lexical character components with their lexical neighbors, it was quantified subjectively through a pretest. A measure of lexical similarity, dubbed the neighbor score, was then used as a predictor, along with sensorimotor variables, for acceptability judgments collected in Experiment 1. Finally, all of these variables, including the acceptability judgments from Experiment 1, were used to predict responses in a same-different task in Experiment 2 and a handwritten copying task in Experiment 3.

While this study is exploratory, the above discussion suggests a number of reasonable expectations about the results: acceptability judgments for non-lexical glyphs should be affected by both lexical and sensorimotor factors, since they reflect the ease of glyph processing; perception-associated variables should affect the perception task more than the production task while production-associated variables should show the reverse pattern; acceptability judgments should affect both perception and production even with lexical and processing-specific variables factored out, since they reflect glyph processing as a whole, including as-yet unknown influences.

2. Pretest: Lexical neighbor judgments

The pretest collected subjective lexical neighbors and similarity judgments for the test items in order to derive a measure of lexical similarity, called here the neighbor score, that could be used to predict responses in the acceptability, perception, and production experiments.

2.1 Method

2.1.1 Participants

23 native Mandarin-speaking traditional character readers at a university in Taiwan were paid a nominal fee for their participation. None reported impairments in language, vision, hearing, or learning.

2.1.2 Materials

The 140 test items consisted of Kitan glyphs that were distinct from any real traditional or simplified Chinese character and contained nothing but contacting strokes, in order to make it easier to quantify the proportion of T-shaped stroke combinations. They were selected from two Kitan databases maintained by Andrew West (West, 2025), with 32 taken from the large script (from Liu & Wang, 2004) and 108 from the small script (from Kane, 2009), respectively displayed in the fonts BabelStone Khitan Large Glyphs and BabelStone Khitan Small Glyphs (an older version of BabelStone Khitan Small Linear) created by Andrew West in the same modern Song style. Four items were set aside as practice items, each pair containing one Chinese-like and one non-Chinese-like form (according to fluent Chinese-reading lab assistants), and the remaining 136 items were used as experimental items.

Sample items are shown in Table 1, along with the lexical neighbor most commonly suggested in the pretest. Summary statistics for the experimental items are shown in Table 2. The acceptance rates come from Experiment 1, and the neighbor scores were computed from the results of the pretest as combined with the results from Experiment 1 (namely, of all the lexical variables computed in the pretest, this is the one that best predicted the acceptability judgments, as is explained in the results section for Experiment 1). Stroke complexity counted the number of line segments (stroke axes) within each stroke, including hooks (following Bohn, 1998); for example, the mean stroke complexity of 女 would be (2+1+1)/3. Cardinality, or the proportion of cardinal (horizontal or vertical) strokes, was computed by first classifying each stroke as containing only cardinal axes (1) or not (0), summing these values, and then dividing by the total; for example, the proportion of cardinal strokes in 女 would be (0+0+1)/3. The T-shape score was computed by counting the number of stroke pairs in which the second extends from the first (assuming rightward/downward stroke directions), subtracting from this the number of stroke pairs in which the first ends at the second, and dividing by the total number of strokes; for example, 下 *xià* 'under' and 上 *shàng* 'over' would have the T-shape scores (2-0)/3 and (1-1)/3, respectively. Confirming what has been found in other writing systems, one-sample *t* tests showed that the Kitan glyphs in the sample significantly favored cardinal axes ($t(135) = 9.65$, $p < .0001$) and T-shaped stroke combinations ($t(135) = 4.43$, $p < .0001$).

Table 1: Sample test items and the lexical neighbors most commonly suggested in the pretest.

	a	b	c	d	e	f	g	h	i
Test item									
Neighbor	美	冊	由	万	木	牟	女	止	及

Table 2: Summary statistics for the test items used in all experiments.

	Min	Max	Mean	SD
Acceptance rate	.056	.917	.429	.224
Neighbor score	2.773	6.478	4.404	0.807
Proportion of cardinal strokes	0	1	.704	.246
T-shape score	-1	0.750	0.120	0.315
Number of strokes	1	9	4.110	1.407
Mean stroke complexity	1	4	1.502	0.445

Notes. Neighbor scores come from the pretest and acceptance rates come from Experiment 1.

2.1.3 Procedure

For each participant, a paper form was prepared by separately randomizing the order of the practice and experimental items in an Excel sheet (different random orders for each participant), and then printing it in 24-point font, with four practice items on the first A4 page and up to 21 items on each of the remaining seven pages. To the right of each item was a blank space for participants to handwrite the real character that came to mind when they saw the test item, whether a traditional character, simplified character (not used in Taiwan but encountered online), Japanese kanji, Cantonese character, or the character-like phonetic notation used in Taiwan (Bopomofo), emphasizing that this character should be the one that seems most similar to the test item. They were also asked to write the pronunciation of the suggested character in phonetic notation (to help the lab assistants identify any unusual or unclearly written characters). Finally, participants were asked to circle one of a row of seven printed digits indicating how similar they thought the fake test item and the suggested real character seemed (1 = very dissimilar, 7 = very similar). There was no time limit, and the whole pretest took around 20 minutes.

2.2 Results

None of the 3,128 trials (23 participants × 136 test items) were skipped or left incomplete. The lexical neighbors suggested by participants were first compiled for each item; examples are shown in Table 1 (see the supplementary materials for the full set). A variety of plausible lexical similarity variables were derived from the participant-suggested neighbors and their accompanying similarity scores, as listed in Table 3. Each of these variables was calculated with and without weighting by log character token frequency (from Tsai, 2006). The first principal component computed across all 12 of these variables for the 136 test items captured 49.6% of the total variance, with strong correlations with the majority of them.

Table 3: Quantifications of lexical similarity and their correlations with the first principal component.

	Frequency weighting	
Variable	Without	With
Mean similarity rating	**.819**	.869
Mean stroke edit distance[a] to all neighbors	-.854	-.847
Most-suggested neighbors	.819[b]	.356[c]
Number of suggested neighbors	-.817	-.341
Mean similarity score for most-suggested neighbor	.668	.693
Mean stroke edit distance[a] to most-suggested neighbor	-.558	-.513

Notes. [a]Number of stroke differences between test items and neighbors assuming fixed stroke orders as judged by fluent Chinese-writing lab assistants. [b]Proportion of the most-suggested neighbor among all those suggested by different participants, reflecting a higher degree of consensus. [c]Log frequency of the most-suggested neighbor. The bolded variable is the one called “neighbor score” in the rest of the paper.

2.3 Discussion

The primary purpose of the pretest was to quantify lexical similarity, but some initial observations can already be made, starting with the stimuli themselves. Even though they were selected without any hypothesis in mind, the sample still conformed to two universal generalizations about orthographic form (cardinal axes were more common than oblique axes, and T-shaped stroke combinations were more common than the inverse), as well as a number of previously observed generalizations about Chinese script in particular. Only two stroke types in the stimuli are not found in modern Chinese: ↪, which appears in only one item, and Z, which appears in five items but is very similar to a complex stroke used in the fluent handwritten form of the common character 之 (*zhī* ‘possessive marker’; only in very careful writing is the Z-shaped portion split into two strokes: ㇇ 一). Moreover, in the stimuli curved strokes and right-hooked vertical strokes only appear on the left, left-hooked vertical strokes never appear on the left, and the lowest of a stack of horizontal strokes is generally the longest, consistent with the Chinese script generalizations of Wang (1983) and Myers

(2019). The items in Table 1 also all have quite obvious Chinese neighbors. Nevertheless, despite the universal influences, Chinese-specific regularities, and lexical neighbors, Experiment 1 will show that the sample items in this table differ greatly in acceptability, as fluent Chinese readers of this paper can probably confirm for themselves (they are arranged here left-to-right from most to least acceptable).

As for the lexical similarity judgment results, the fact that the first principal component captured so much of the variance of so many different quantifications of subjective lexical similarity suggests that this concept has some core psychological validity. It is also striking that character frequency mattered so little. This can be seen not only in the correlations in Table 3, but also in the suggested lexical neighbors in Table 1, which include relative rarities (牟 *móu* 'cow's moo' for item (f)) and even characters outside the traditional system used by the Taiwanese participants (万 for item (d) is the simplified form of 萬 *wàn* 'ten thousand'). Lexicality similarity here thus seems to reflect knowledge of the character inventory, not just the ease of lexical access (see Bybee, 2007, for the linguistic implications of type vs. token frequency).

3. Experiment 1: Acceptability judgments

This experiment collected acceptability judgments to determine which of the quantifications of lexical similarity best predicted them, along with variables associated with perceptual and motoric processing. The by-item acceptability rates could then also be used as a predictor in the perception and production experiments.

3.1 Method

3.1.1 Participants

A new group of 36 participants from the same population as the pretest were paid a nominal fee for their participation. None reported impairments in language, vision, hearing, or learning.

3.1.2 Materials

The stimuli were identical to those used in the pretest.

3.1.3 Procedure

Each participant was presented with a randomly ordered series of items using PsychoPy (Peirce, Hirst, & MacAskill, 2022), divided into a practice session with four items and the main experiment with 136 items. Participants were asked to decide if each test item was structured like or unlike Chinese characters (see supplementary materials for the original Chinese instructions). Binary responses indicated acceptance or rejection of Chinese-likeness, with the ‘S’ key on the left of the computer keyboard representing rejection and the ‘L’ key on the right representing acceptance. Each trial began with a fixation point (‘o’) for one second, followed by 500 ms of a blank screen, and then the item was displayed for two seconds, in black on a white background within a 8 cm^2 square, after which it disappeared. The trial ended when a response was made or else four seconds after display onset, and before the next trial began there was a blank screen for 500 ms. The use of binary responses allowed for the collection of reaction times, and also made the task quite easy to perform, with participants finishing well within ten minutes.

3.2 Results

Participants skipped 46 of the 4,896 trials (36 participants × 136 test items), with no individual skipping more than 9%. Dropping implausibly fast responses (shorter than 100 ms) left 4,833 trials for analysis. Using the lme4 package (Bates, Maechler, Bolker, & Walker, 2015) in R (R Core Team, 2025), separate maximal mixed-effects logistic regression models were created to predict the binary judgment responses from each of the twelve quantifications of lexical similarity derived from the pretest results (z-scored) and from the first principal component (see Table 3). The best of these 13 models, according to Akaike’s information criterion and an analog of R^2 for mixed-effects logistic regression (Nakagawa, Johnson, & Schielzeth, 2017) as implemented in R’s MuMIn package (Bartoń, 2024), was the model predicting responses from the mean neighbor-target similarity rating on the seven-point scale (1 = very dissimilar, 7 = very similar) without weighting by character frequency (bolded cells in Table 3). This variable was thus redubbed the neighbor score, and it is used as the sole quantification of lexical similarity in all of the remaining analyses.

As shown in Table 2, the mean acceptance rate was .429, suggesting only a slightly greater tendency to reject rather than accept items. The left plot in Figure 1 also shows that the distribution of by-item acceptance rates was somewhat bimodal, unlike the monomodal distribution of the neighbor scores. The two variables were correlated ($r(134) = .619$, $p < .001$), although as suggested by the coefficient of determination ($R^2 = .383$) and the right plot in Figure 1, most of the variance in acceptance rates was unexplained by neighbor scores (nor by the other fixed variables, as demonstrated shortly).

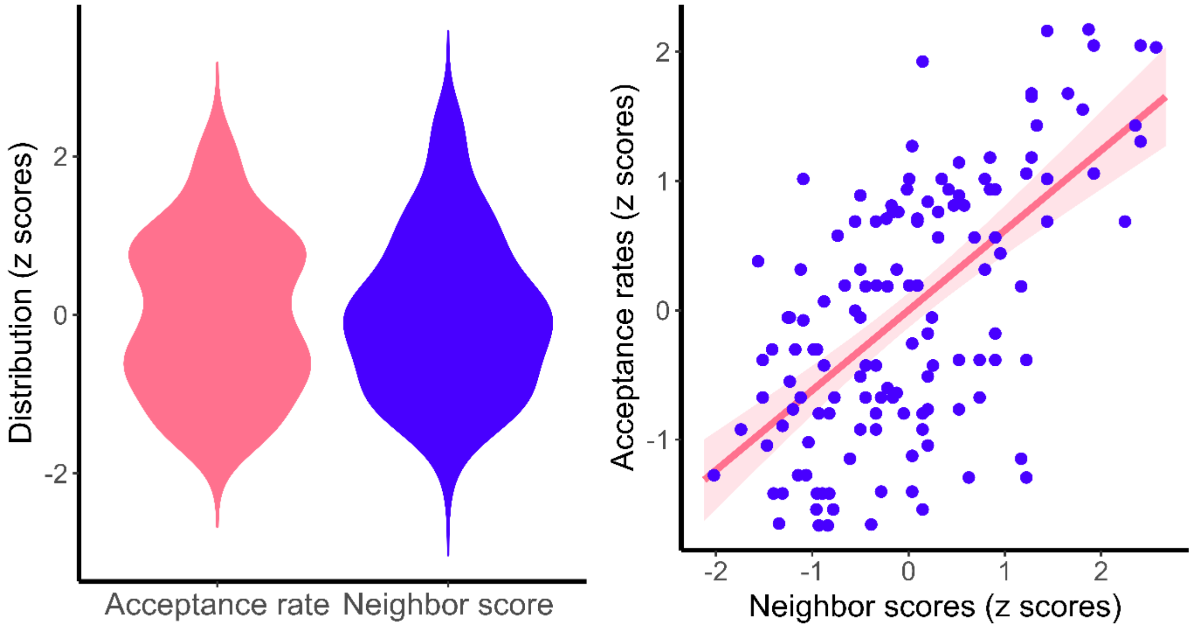

Figure 1: Distribution shapes (left) and correlation (right) of by-item neighbor scores and acceptance rates (both *z*-scored); the error band represents the 95% confidence interval.

The acceptability judgment responses were then modeled using mixed-effects logistic regression from the neighbor scores (lexical similarity), the proportion of cardinal strokes (expected to reflect visual processing), the T-shape score (expected to reflect motoric processing), and the number of strokes and mean stroke complexity (both expected to affect both visual and motoric processing). Only the last variable had to be log-normed, but all predictors were *z*-scored to improve convergence and allow for the comparison of effect sizes. There was no collinearity (all VIF < 1.9). To ensure convergence in a maximal model with random intercepts for both participants and items and random slopes for participants (Barr, Levy, Scheepers, & Tily, 2013), the model was fit using weakly informative Bayesian priors with R's blme package (Chung, Rabe-Hesketh, Dorie, Gelman, & Liu, 2013).

The model found that acceptability was significantly increased by neighbor scores (B = 0.897, SE = 0.113, z = 7.952, $p < .0001$) and by the number of strokes (B = 0.524, SE = 0.129, z = 4.072, $p < .0001$); no other effects reached significance (all $p > .1$). The model as a whole had an estimated R^2 (via R's MuMIn package) of only up to 15.4% for all of the fixed variables (up to 53.1% if the random variables were also included), consistent with the right plot in Figure 1 that suggested that most of the variance in acceptability remained unexplained by the tested variables.

To better understand how these judgements were made in real time, a linear mixed-effects model was used to predict reaction time (log-normed) from response choice (accept vs. reject, effect-coded to allow for the testing of interactions with it), the variables above, and the interactions between them and response choice; dependent and independent variables were all *z*-scored. Because model convergence again required weakly informative Bayesian priors, where degrees of freedom are difficult to define, *p*-values were estimated using Type II Wald chi-square tests (via R's car package: Fox & Weisberg, 2019), although for completeness, the original model's *t* values are reported as well. The analysis is shown in Table 4.

Table 4: Linear mixed-effects model of reaction times in Experiment 1.

	β	*SE*	*t*	χ^2	*p*
Intercept	0.103	0.077	1.339		
Response	0.063	0.043	1.482	3.899	.048
Neighbor score	-0.068	0.020	-3.331	14.850	< .001
Cardinality	-0.013	0.018	-0.731	0.133	.715
T-ness	0.010	0.017	0.601	0.096	.756
Stroke number	0.039	0.022	1.793	3.340	.068
Stroke complexity	0.046	0.021	2.207	4.823	.028
Response × neighbor score	-0.061	0.020	-3.096	9.583	.002
Response × cardinality	-0.039	0.018	-2.098	4.402	.036
Response × T-ness	0.021	0.016	1.313	1.723	.189
Response × stroke number	-0.034	0.026	-1.305	1.704	.192
Response × stroke complexity	0.005	0.023	0.232	0.054	.817

The only unambiguously significant effects ($|t| > 2$ and $p < .05$) were neighbor scores and stroke complexity, which sped and slowed responses respectively, and the interactions of response choice with neighbor score and cardinality: neighbor scores sped up acceptances but had no effect on rejections (Figure 2), while cardinality sped acceptances and slowed rejections (Figure 3). Overall there was also a slowing of acceptances relative to rejections, though this was only marginally significant ($|t| < 2$, $p < .05$).

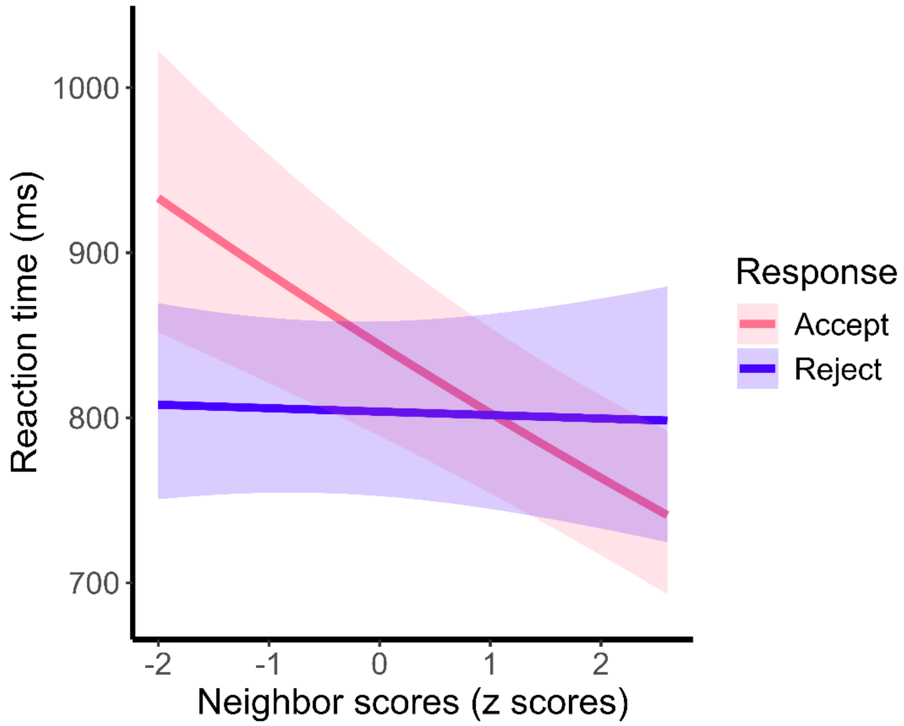

Figure 2: The effect on reaction times of response (accept vs. reject), as modulated by neighbor scores; the error bands represent 95% confidence intervals.

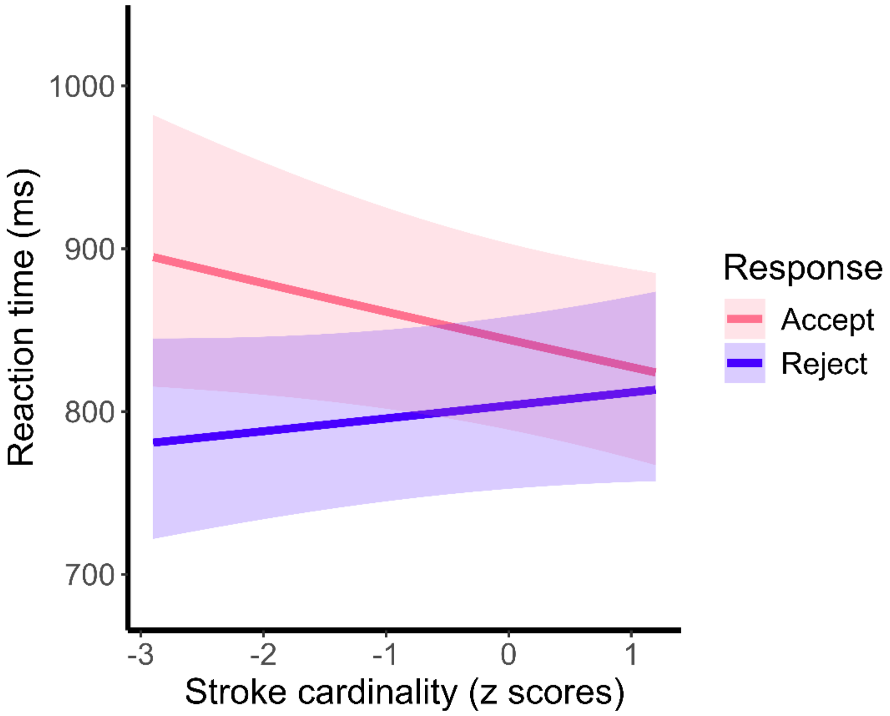

Figure 3: The effect on reaction times of response (accept vs. reject), as modulated by stroke cardinality; the error bands represent 95% confidence intervals.

3.3 Discussion

Similar to what has repeatedly been found in spoken language experiments, acceptance for non-lexical characters was greater the higher the degree of lexical similarity, quantified here with subjective neighbor scores. Surprisingly, acceptability was also higher for items with a greater number of strokes. This preference may simply be due to a combinatorial artifact: the fewer the strokes, the odder they, their contacts, or positions must be in order for the item not to be lexical. More intriguingly, may also be that the Taiwanese participants, fluent in traditional Chinese orthography, preferred a many-stroke "look" to the sparser appearance of "out-group" systems like simplified characters or Roman letters. In either case, this finding already suggests that glyph acceptability cannot be derived in a simple way from simple processing variables like visual complexity. None of the other variables that had been expected to affect perception or production showed significant effects ($p > .1$). Moreover, unlike the neighbor scores, the distribution of by-item acceptance rates was somewhat bimodal, as if in the participants' minds, items divided roughly into possible and impossible characters. This pattern should not be ascribed to the use of a binary response scale: not only is the binomial distribution itself not bimodal, but phonotactic studies have also observed a tendency towards categorical judgments in acceptability tasks even when using numerical response scales (Gorman, 2013).

The reaction time results also shed some light on how acceptability judgments are made. Acceptances tended to be slower to make than rejections and were also significantly faster for items more similar to lexical components, whereas the reaction times for rejections were unaffected by lexical similarity. This pattern suggests that rejections were made on the basis of analyses that did not depend on lexical access. Consistent with this interpretation, stroke cardinality not only sped acceptances but also slowed rejections, because as a visual variable it affected the earlier stages when rejections were made, though its influence on the acceptances shows that it remained relevant throughout the whole decision-making process. As expected, then, acceptability judgments are multifaceted, being sensitive to both lexical and perceptual factors, even for non-lexical orthographic materials. It is not clear how participants set the threshold between the rejection and acceptance modes. Huang and Ferreira (2020) suggest applying signal detection theory to acceptability judgments, but their approach could not be used here because it requires the a priori identification of "correct" and "incorrect" responses, precisely what this experiment was intended to determine empirically.

4. Experiment 2: Same-different task

The primary goal of this experiment was to test whether acceptability helps predict the real-time perception of non-lexical Chinese characters, even after factoring out the influences of lexical similarity and universal sensorimotor forces.

4.1 Method

4.1.1 Participants

A new group of 36 participants from the same population as the pretest and Experiment 1 were paid a nominal fee for their participation. None reported impairments in language, vision, hearing, or learning.

4.1.2 Materials

The same image files as in Experiment 1 were presented in pairs, half with the same character twice and the other half with two different characters. "Different" pairs randomly combined character images (separately for the practice and experimental trials) so that each appeared once in each position. There were thus eight practice pairs (four each of "same" and "different") and 272 experimental pairs (136 of each type of pair). All participants received the same set of pairs.

4.1.3 Procedure

Participants were presented with a series of randomly ordered trials using PsychoPy. The characters had the same size as in Experiment 1, and were presented in side-by-side pairs centered in the middle of the screen. Participants were asked to decide if the characters in each pair were the same or different by pressing the 'L' key on the right or 'S' key on the left, respectively. To encourage spontaneous reactions, only up to three seconds were allowed for responses (cf. the four seconds allowed for Experiment 1). Otherwise the trials were structured identically to Experiment 1. The experiment took around 12 minutes to complete.

4.2 Results

Of the 9,792 total trials (36 participants × 136 items × 2 pair types), only 28 were skipped, with no participant skipping more than 2%. All participants had an accuracy rate over 93% (within 1 SD of the mean accuracy of 98%), and no response was implausibly fast (shorter than 100 ms), so there were 9,764 valid trials for analysis. A maximal linear mixed-effects model showed no significant effect of accuracy on log response latencies (*p* > .3), so there was no speed-accuracy trade-off. The remaining analyses focused just on the 4,878 trials with same-character pairs, because the "different" pairs contained two distinct stimuli, making it impossible to predict responses to them from item-level variables. The independent variables were again neighbor scores, proportion of cardinal strokes, T-ness scores, number of strokes and log mean stroke complexity, supplemented now with the by-item acceptance rates from Experiment 1, all *z*-scored; there was no collinearity (all VIF < 2.2).

A maximal mixed-effects logistic regression found no effect of any of these predictors on response accuracy (all *p* > .3), presumably because mean accuracy was near ceiling (98%). However, there were richer results from a maximal linear mixed-effects model on log-normed reaction times to the 4,779 trials with accurate responses to the "same" pairs. Because model convergence again required the use of weakly informative priors, the *p* values were estimated using Type II Wald chi-square tests, and as for Experiment 1, the original model's *t* values are also reported. As shown in Table 5, the effects with unambiguous statistical significance (|*t*| > 2 and *p* < .05) were the proportion of cardinal strokes and the number of strokes, the former speeding up responses and the latter slowing them down. Acceptance rate also sped up responses marginally (|*t*| = 1.963, .1 > *p* > .05). T-shape scores and mean stroke complexity did not have significant effects (both *p* > .1), nor did neighbor scores (*p* > .9).

Table 5 Linear mixed-effects model of reaction times for correct "same" responses in Experiment 2

	β	*SE*	*t*	χ^2	*p*
Intercept	< 0.001	0.101	0.002		
Acceptance rate	-0.059	0.030	-1.963	3.854	.050
Neighbor score	-0.003	0.029	-0.090	0.008	.929
Cardinality	-0.052	0.022	-2.366	5.599	.018
T-ness	0.032	0.023	1.394	1.943	.163
Stroke number	0.116	0.032	3.575	12.783	< .001
Stroke complexity	0.033	0.028	1.152	1.327	.249

4.3 Discussion

While accuracy was too close to ceiling to reveal any effects of the predictors, correct responses to "same" pairs were significantly faster for items with a higher proportion of cardinal axes, as would be expected based on the arguments from Changizi and Shimojo (2005) and Morin (2018) that writing systems favor such strokes for visual reasons. Responses were also significantly slower for items with a greater number of strokes, which increases visual complexity. Neither mean stroke complexity nor T-shape scores seemed to matter. Notably, given the goals of this study, responses were also marginally faster for items with higher acceptance rates from Experiment 1 ($p = .05$), while the neighbor scores had no effect at all ($p > .9$), even though none of the predictors are collinear. In other words, even in a perceptual task that did not ask participants to consider acceptability, they still seemed to take acceptability into account, and apparently in a way that did not depend on comparisons with lexical neighbors nor on any of the sensorimotor variables included in the model.

5. Experiment 3: Handwriting

The goals of this final experiment were to test whether acceptability affects handwriting as well, even after factoring out lexical and universal sensorimotor influences. Because the items are non-lexical, handwriting could not be elicited via non-written stimuli, as is done, for example, in the handwriting megastudy by Wang et al. (2020), where characters were prompted by disambiguating spoken phrases. Instead, the experiment here used a character copying task, the handwriting equivalent of the repetition naming task. Unlike a “pure” handwriting task, character copying requires participants to perceive the stimuli, store the image in working memory, generate articulatory plans, and finally implement the handwriting motor programs, thereby engaging both central and peripheral processes (Van Galen, 1991). Response latencies (time to initiate writing) may thus reflect any of these processes, whereas writing duration (from the start to the end of writing) should be affected primarily by motor implementation (peripheral processing).

5.1 Method

5.1.1 Participants

A new group of 43 participants from the same population as the previous experiments were paid a nominal fee for their participation. None reported impairments in language, vision, hearing, or learning. After dropping 11 participants for skipping trials or high error rates (see results), there were 32 participants for analysis, three of whom were left-handed (including them did not affect the results, as demonstrated below).

5.1.2 Materials

The stimuli were the same non-lexical characters as those used in all of the previous experiments, except that to ease perception and to reduce superficial similarity between the stimuli and handwritten forms, they were presented in white on a black background.

5.1.3 Procedure

A Wacom® One tablet and stylus, controlled by PsychoPy on a separate computer, was used to display instructions and stimuli and to collect the writing data. Each participant was presented with a randomly ordered series of trials, with four practice trials and 136 experimental trials. Participants were told that they would see a series of unfamiliar characters and were asked to write each on the tablet within a 5 cm^2 writing area in the center of the tablet screen. Trials began with a fixation point for 1 s (‘o’), followed by a 500 ms blank screen. The stimulus then appeared in the writing area for 500 ms (long enough to view clearly but not long enough to study carefully), after which it was replaced with a visual mask for 100 ms (white ‘X’s on a black background) to prevent participants from attempting to trace the characters from retinal afterimages. The writing area then became blank, after which participants wrote their response within it using the stylus, making marks in black on a white background. Trials ended only after the participant tapped the stylus on a button labeled ‘OK’ at the bottom of the screen. The next trial then started immediately. The whole experiment took around 20 minutes.

5.2 Results

R scripts (see Author, 2023, for details) compiled the raw data (x and y coordinates, time, and stylus contact status every 20 ms), determined whether the whole character was written within the writing area, and calculated the response latency (time in ms from the appearance of the item to the first stylus contact within the writing area), writing duration (time in ms from the onset of writing to the end), and the number of handwritten strokes (based on the number of changes in pen contact within the writing area). The total number of trials was 5,848 (43 participants × 136 items), but two participants who had more than 10% invalid trials (by skipping or writing outside the writing area) were dropped; 100 invalid trials from the remaining participants were also dropped. No response latencies were less than 100 ms from the appearance of the writing area (i.e., faster than 700 ms from stimulus onset). Correct responses were defined as those sufficiently visually similar to the test items, as agreed by a majority of five fluent Chinese-reading lab assistants, while also containing the same number of strokes as calculated automatically, with target stroke number determined by the lab assistants. Because of the difficulty of the task (rapidly copying briefly displayed non-lexical characters), the accuracy threshold was set to a generous 70%, which led to the dropping of nine additional participants below this threshold. This left an overall accuracy rate of 78.6% and a final tally of 32 participants and 4,304 trials for analysis.

A maximal linear mixed-effects model (with p values computed from the t values using Satterthwaite's method) found that (log) response latencies were significantly faster for correct responses than for incorrect ones ($B = -0.03$, $SE = 0.009$, $t(48.46) = -3.4$, $p = .001$), confirming that there was no speed-accuracy trade-off. A maximal mixed-effects logistic regression model (with weakly informative priors to ensure convergence) was used to predict accuracy from acceptance rates, neighbor scores, proportion of cardinal strokes, T-ness scores, number of strokes, and log mean stroke complexity (with no collinearity: all VIF < 2.2); all independent variables were z-scored, and p values were computed using Wald tests. As shown in Table 6, this model found that accuracy was significantly higher for more acceptable items and for items with a higher proportion of cardinal strokes, and lower for items with a greater total number of strokes; no other effects reached significance, including neighbor scores ($p > .6$).

Table 6 Mixed-effects logistic regression model of accuracy in experiment 3

	B	*SE*	*z*	*p*
Intercept	1.703	0.116	14.676	< .001
Acceptance rate	0.596	0.155	3.839	< .001
Neighbor score	-0.070	0.140	-0.498	.619
Cardinality	0.476	0.111	4.287	< .001
T-ness	0.143	0.108	1.331	.183
Stroke number	-0.737	0.152	-4.853	< .001
Stroke complexity	-0.124	0.139	-0.893	.372

A maximal linear mixed-effects model (with weakly informative priors and p values estimated from Wald chi-square tests) was then used to predict z-scored log response latencies for correctly written items from the same z-scored independent variables. As shown in Table 7, the same three effects were the only ones to achieve unambiguous statistical significance ($|t| > 2$ and $p < .05$): reaction times were faster for more acceptable items and for items with a higher proportion of cardinal strokes, and slower for items with more strokes.

Log mean stroke complexity showed a marginal trend toward slowing response latencies ($|t| > 1.7$, $.1 > p > .05$), and once again, neighbor score showed no effect ($p > .1$).

Table 7 Linear mixed-effects model of response latencies for correctly written characters in Experiment 3

	β	*SE*	*t*	χ^2	*p*
Intercept	-0.001	0.110	-0.006		
Acceptance rate	-0.124	0.035	-3.596	12.931	< .001
Neighbor score	0.043	0.030	1.456	2.119	.145
Cardinality	-0.069	0.022	-3.087	9.532	.002
T-ness	-0.013	0.021	-0.629	0.396	.529
Stroke number	0.159	0.034	4.633	21.461	< .001
Stroke complexity	0.050	0.028	1.787	3.192	.074

Finally, a linear mixed-effects model was used to predict *z*-scored log durations from the same variables, this time dropping correlations between random intercepts and slopes due to convergence problems unfixable even with the quasi-Bayesian approach; *p* values were estimated using Satterthwaite's method. As shown in Table 8, this model found that duration significantly increased in items with more strokes and with more complex strokes. No other effects reached significance (all $p > .1$).

Table 8 Linear mixed-effects model of writing durations for correctly written characters in Experiment 3

	β	*SE*	*df*	*t*	*p*
Intercept	0.008	0.111	35.163	0.075	.940
Acceptance rate	-0.061	0.042	126.997	-1.449	.150
Neighbor score	0.012	0.039	123.779	0.296	.768
Cardinality	0.045	0.032	137.741	1.414	.160
T-ness	0.036	0.030	126.388	1.173	.243
Stroke number	0.640	0.042	132.406	15.221	< .001
Stroke complexity	0.092	0.041	132.184	2.258	.026

All of these analyses remained virtually identical when the three left-handed participants were dropped, with the same coefficient signs and significance levels. Because of the extremely small number of left-handers, handedness could not be included as a factor in any analysis.

5.3 Discussion

As shown by both accuracy and response latencies, items with a greater number of strokes (and, marginally, more complex strokes) were harder to prepare for handwriting, but those with a higher proportion of cardinal strokes were easier; once again, T-ness had no effect. Given the nature of the copying task, these effects could have occurred during perception, rather than during production planning. If Changizi and Shimojo (2005) and Morin (2018) are right, the effect of stroke cardinality must have been perceptual, but the number and complexity of strokes should also be able to affect production planning. Importantly for this study's goals, responses also benefited from acceptability, with more acceptable items being responded to more accurately and more quickly, even though, as in Experiment 2, neighbor scores had no effect on either measure. Even though the effect of acceptability on response latencies for writing seemed to be stronger in the handwriting task in Experiment 3 (Table 7: $\beta = -0.124$, $p < .001$) than in the perceptual task in Experiment 2 (Table 5: $\beta = -0.059$, $p = .05$), a mixed-effects linear regression on response latencies across both experiments (with weakly informative priors and p values estimated from Wald chi-square tests) found no significant interaction between acceptability and task ($\beta < 0.0001$, $SE = 0.013$, $t = 0.007$, $\chi^2 < 0.0001$, $p > .9$). Only one interaction with task approached significance, namely a marginal interaction involving T-ness ($\beta = 0.0187$, $SE = 0.011$, $t = 1.695$, $\chi^2 = 2.873$, $p = .09$), which was inhibitory in Experiment 2 but had no effect in Experiment 3. However, because this variable did not show even marginal effects within any single task, and should be associated with production rather than perception, it seems unlikely that this particular observation will be replicable.

Writing duration showed no effect of stroke cardinality, as expected if this variable affects perception but not production, while durations were, unsurprisingly, longer for items with more strokes and with more complex strokes. There was also no effect of T-shape proportion, despite this type of stroke combination being preferred not just in line drawing experiments but within the stimuli themselves. It may be that this particular measure is not just not sensitive enough to pick up on real-time motoric constraints on writing, at least in a stimulus set already dominated by T-shapes. Acceptability also failed to have a significant effect on writing duration, suggesting that motor control, as opposed to perception and production planning, may not be directly influenced by well-formedness, though as with all null results, further research is needed to clarify this point.

6. General discussion

This study examined the implicit productive knowledge that Chinese readers and writers have about the form of character components by testing them in a variety of tasks with non-lexical components. The results are summarized in Table 9, which shows the facilitative (+) and inhibitory (-) effects on accuracy, speed (shorter response latency), and completion (shorter duration) as predicted by acceptability, neighbor score (subjective lexical similarity), stroke axis cardinality (expected to affect visual processing), T-ness (a measure of stroke combination expected to affect manual articulation), and stroke number and complexity (expected to affect both visual and motoric processing).

Table 9 Summary of all experimental results

	Experiment 1 (acceptability)		Experiment 2 (same-different)		Experiment 3 (handwriting)		
	Acceptability	Speed	Accuracy	Speed	Accuracy	Speed	Completion
Acceptability	NA	NA		(+)	+	+	
Neighbor score	+	+					
Cardinality		+		+	+	+	
T-ness							
Stroke number	+			-	-	-	-
Stroke complexity						(-)	-

Notes. Speed: shorter response latency; Completion: shorter duration; NA: not applicable; +/-: facilitation/inhibition; (): marginal (.1 > *p* > .05).

The results confirmed the expected influences of perceptual variables (cardinality, stroke number, and stroke complexity) and to a lesser extent articulatory variables (stroke number and stroke complexity, but not T-ness) on real-time processing. Unexpectedly, however, there were no significant differences in their effects across the perception and production tasks. This may be because the production task involved the copying of visual stimuli (the non-lexical stimuli could not be prompted in any other way), and thus it necessarily engaged perceptual and working memory processes in addition to motor planning. Acceptability judgments were also positively influenced not just by lexical similarity but also by one of the perceptual factors (higher cardinality sped responses) and, surprisingly, by stroke number as well (acceptability was higher with more strokes). This stroke number effect could not have been sensorimotor because stroke number effects were inhibitory in the real-time processing tasks, and so must have reflected higher-level knowledge, perhaps a learned preference among traditional character readers for greater visual complexity.

One of the most important of this study's theoretical motivations was that acceptability judgments for non-lexical Chinese character components should reflect overall processing rather than any specific process, as with form-based acceptability judgments more generally (Avcu et al., 2023; Frisch & Zawaydeh, 2001; Gouskova & Becker, 2013). The results did indeed suggest that acceptability was irreducible to any other variable, including lexical similarity, the variable that was its strongest correlate. This can be seen in a number of independent observations: by-item neighbor scores accounted for only 38.3% of the variance in by-item acceptance rates, acceptance rates formed a bimodal distribution while neighbor scores did not, neighbor scores only affected response latencies when accepting items but not when rejecting them, and acceptance rates affected the same-different and handwriting tasks while neighbor scores did not, despite the lack of collinearity. Acceptability effects also went in the direction one would expect if they reflect ease of processing (Gross, 2020; Topolinski & Strack, 2009), speeding and/or increasing the accuracy of both perception and production tasks.

Nevertheless, drawing a line between acceptability and lexical similarity remains difficult, for instructive reasons. It may seem that for basic orthographic glyphs, lexical similarity could be defined objectively by adopting a distance metric like those long used in optical character recognition (OCR) algorithms. The core problem is that while a variety of such algorithms have been proposed (Vynckier, 2025), the only "correct" one, from a cognitive perspective, is what humans actually use, and the only way to find it is with the help of subjective judgments. Current off-the-shelf OCR tools, like those in Adobe Acrobat or Tesseract (the latter tested via its R implementation: Ooms, 2025), both of which produce highly accurate results for identifying traditional Chinese characters, tend to offer Chinese neighbors quite different from those suggested by this study's pretest participants. For example, item (a) in Table 1, which the pretest participants found quite similar to 美 (*měi* 'beautiful'), was identified by Acrobat as 芙 (*fú* 'lotus'); the human participants seemed to care more about matching stroke type than pixel-by-pixel identity (the latter character is also rarer than the former, but the human-suggested neighbor 牟 for item (f) is also relatively rare and the simplified 万 for item (d) is outside the traditional system entirely). Tesseract performs even worse than Acrobat, presumably because it was trained on continuous text rather than isolated characters. A deep learning network could probably be trained to mimic human orthographic neighbor judgments, but the deeper such networks would need to become to adequately fit the data, the less they would quantify superficial lexical similarity, and the more they would begin to simulate the complex and multifaceted system that underlies acceptability judgments.

If neither lexical similarity nor sensorimotor variables suffice to explain acceptability judgments, what does? As noted in the pretest discussion, the stimuli in this study already conformed to most of the known generalizations regarding Chinese component structure, both universal (preferences for cardinal axes and T-shaped stroke combinations) and script-specific (the use of lexical Chinese strokes, with curved, hooked, and lengthened strokes in their allowed positions), so productive knowledge of component form must somehow go beyond these. While T-ness did not seem to matter as much as expected, other constraints on stroke combinations are worth exploring, not just universal metrics of juncture complexity (Altmann, 2005) but also script-specific constraints. For example, item (h) in Table 1 uses the Chinese stroke ㇕ in a non-Chinese way, since in lexical components this stroke always makes stroke contact on the left, as in 片 (*piàn* ‘slice’). Unfortunately, it is not obvious how such constraints could be reduced to simple principles; note that the otherwise very similar stroke ㇆ does not obey this constraint, as in 司 (*sī* ‘manage’).

The collection and analyses of such observations could be made more systematic with the development of more detailed theories of glyph representation. One interesting approach may be to adapt the traditional phonological division between content (segments and features) and structure (prosody, like syllables and stress feet). Primus (2004) and Watt (1988) decompose individual Roman letters into strokes, each represented by features, while stroke combinations are handled in a separate part of their theories, morphology-like in Watt (1988) and prosody-like in Primus (2004). Chinese character strokes have also been given featural analyses by Myers (2019), Peng (2017), and Wang (1983), with Myers (2019) explicitly describing constraints on their combinations as analogous to prosody. Kim et al. (2025) provide empirical support for orthographic features as well, but they apply them to entire glyphs; when they find that more distinctive features are required for glyph inventories than for phoneme inventories, it may be because they are conflating what are actually distinct representational levels.

Regardless of the theoretical direction one might ultimately want to follow, the results reported here demonstrate that fluent Chinese readers and writers have productive knowledge of the well-formedness of individual character components, not just of component combinations, that this knowledge goes beyond a variety of orthographic generalizations currently recognized in the literature, both universal and script-specific, that it is not reducible to subjective lexical similarity nor to a variety of sensorimotor variables, and finally that it is nevertheless reflected, in some way, in both perception and production. If these results seem to raise more questions than they answer, it is due to the curious neglect of glyph form processing as a research topic, whether in psycholinguistics, computational modeling, or cognitive science more generally. It is thus hoped that future research will attempt to replicate and extend this study within the Chinese script, both traditional and simplified, build on the approach illustrated here in the study of other writing systems and in artificial orthography experiments, and to explore the nature of this purely form-based orthographic knowledge in greater detail, including its similarities with, and differences from, the natural phonology of speech and signing.

Data Accessibility

The full stimulus list, experimental instructions, experimental data, and statistical analyses are available at https://osf.io/x6btj/overview?view_only=f3709e9293b945beabf74f300c24b52e.

References

Altmann, G. (2005). Script complexity. *Glottometrics, 8*, 68-74.

Arendsen, J., van Doorn, A. J., & de Ridder, H. (2010). Acceptability of sign manipulations. *Sign Language & Linguistics*, *13*(2), 101-155. https://doi.org/10.1075/sll.13.2.01are

Avcu, E., Newman, O., Ahlfors, S. P., & Gow Jr, D. W. (2023). Neural evidence suggests phonological acceptability judgments reflect similarity, not constraint evaluation. *Cognition*, *230,* 105322. https://doi.org/10.1016/j.cognition.2022.105322

Baayen, R. H. (2010). A real experiment is a factorial experiment? *The Mental Lexicon*, *5*(1), 149-157. https://doi.org/10.1075/ml.5.1.06baa

Bailey, T. M., & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language, 44*(4), 568-591. https://doi.org/10.1006/jmla.2000.2756

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68* (3), 255-278. https://doi.org/10.1016/j.jml.2012.11.001

Bartoń, K. (2024). MuMIn: Multi-model inference. R package version 1.48.4. https://doi.org/10.32614/CRAN.package.MuMIn

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01

Berent, I., & Lennertz, T. (2010). Universal constraints on the sound structure of language: Phonological or acoustic? *Journal of Experimental Psychology: Human perception and performance*, *36*(1), 212-223. https://doi.org/10.1037/a0017638

Best, C. T. (2020). Perception of non-native click consonant contrasts: Implications for theories of speech perception. In B. Sands (Ed.), *Click consonants* (pp. 179-209). https://doi.org/10.1163/9789004424357_006

Bohn, H. (1998). *Quantitative Untersuchungen der modernen chinesischen Sprache und Schrift*. Hamburg: Kovač.

Bybee, J. (2007). *Frequency of use and the organization of language*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195301571.001.0001

Byrne, B., & Carroll, M. (1989). Learning artificial orthographies: Further evidence of a nonanalytic acquisition procedure. *Memory & Cognition, 17*, 311-317. https://doi.org/10.3758/BF03198469

Chang, L. Y., Tseng, C. C., Perfetti, C. A., & Chen, H. C. (2022). Development and validation of a Chinese pseudo-character/non-character producing system. *Behavior Research Methods, 54*, 632-648. https://doi.org/10.3758/s13428-021-01611-8

Changizi, M.A., & Shimojo, S. (2005). Character complexity and redundancy in writing systems over human history. *Proceedings of the Royal Society of London B: Biological Sciences, 272*(1560), 267-275. https://doi.org/10.1098/rspb.2004.2942

Chen, J.-Y., & Cherng, R.-J. (2013). The proximate unit in Chinese handwritten character production. *Frontiers in Psychology, 4*, 517. https://doi.org/10.3389/fpsyg.2013.00517

Chetail, F. (2015). Reconsidering the role of orthographic redundancy in visual word recognition. *Frontiers in Psychology*, *6*, 645. https://doi.org/10.3389/fpsyg.2015.00645

Chomsky, N. & Halle, M. (1965). Some controversial questions in phonological theory. *Journal of Linguistics, 1* (2), 97-138. https://doi.org/10.1017/S0022226700001134

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., & Liu, J. (2013). A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika, 78*(4), 685-709. https://doi.org/10.1007/s11336-013-9328-2

Cohn, A. C., Fougeron, C., & Huffman, M. K. (2017). Laboratory phonology. In S. J. Hannahs & A. Bosch (Eds.), *The Routledge Handbook of Phonological Theory* (pp. 504-529). London: Routledge. https://doi.org/10.4324/9781315675428-18

Duan, X., & Cai, Z. (2024). Chinese Character Processing. In M. Aronoff (Ed.), *Oxford research encyclopedia of linguistics*. https://doi.org/10.1093/acrefore/9780199384655.013.1012

Dziubalska-Kołaczyk, K. (2014). Explaining phonotactics using NAD. *Language Sciences, 46*, 6-17. https://doi.org/10.1016/j.langsci.2014.06.003

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science, 12*(6), 499-504. https://doi.org/10.1111/1467-9280.00392

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.) Thousand Oaks CA: Sage.

Frisch, S. A., & Zawaydeh, B. A. (2001). The psychological reality of OCP-Place in Arabic. *Language, 77* (1), 91-106. https://doi.org/10.1353/lan.2001.0014

Gong, S., Zhang, J., & Fiorentino, R. (2024). Phonological well-formedness constraints in Mandarin phonotactics: Evidence from lexical decision. *Language and Speech*, *67*(3), 676-691. https://doi.org/10.1177/00238309231182363

Goodnow, J. J., & Levine, R. A. (1973). "The grammar of action": Sequence and syntax in children's copying. *Cognitive Psychology*, *4*(1), 82-98. https://doi.org/10.1016/0010-0285(73)90005-4

Gorman, K. (2013). *Generative phonotactics*. Doctoral dissertation, University of Pennsylvania.

Gouskova, M., & Becker, M. (2013). Nonce words show that Russian yer alternations are governed by the grammar. *Natural Language & Linguistic Theory, 31*(3), 735-765. https://doi.org/10.1007/s11049-013-9197-5

Gross, S. (2020). Linguistic intuitions: Error signals and the Voice of Competence. In S. Schindler, A. Drożdżowicz, & K. Brøcker (Eds.) *Linguistic intuitions: Evidence and method* (pp. 13-32). Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198840558.003.0002

Han, S. J., Kelly, P., Winters, J., & Kemp, C. (2022). Simplification is not dominant in the evolution of Chinese characters. *Open Mind, 6*, 264-279. https://doi.org/10.1162/opmi_a_00064

Hayes, B., & White, J. (2013). Phonological naturalness and phonotactic learning. *Linguistic Inquiry*, *44*(1), 45-75. https://doi.org/10.1162/LING_a_00119

He, X., & Tong, X. (2017). Statistical learning as a key to cracking Chinese orthographic codes. *Scientific Studies of Reading, 21*(1), 60-75. https://doi.org/10.1080/10888438.2016.1243541

Hsieh, C. Y., Lin, W. C., Li, M. F., & Wu, J. T. (2021). The concept of neighborhood underlying the phonetic consistency effect in Chinese character recognition: A four-decade review and beyond. *Reading and Writing, 34*, 1225-1246. https://doi.org/10.1007/s11145-020-10119-0

Huang, Y., & Ferreira, F. (2020). The application of signal detection theory to acceptability judgments. *Frontiers in Psychology, 11*, 73. https://doi.org/10.3389/fpsyg.2020.00073

Hue, C.-W. (2003). Number of characters a college student knows. *Journal of Chinese Linguistics, 31*(2), 300-339.

Jameson, K. (1994). Empirical methods for evaluating generative semiotic models: An application to the roman majuscules. In W.C. Watt (Ed.), *Writing systems and cognition: Perspectives from psychology, physiology, linguistics, and semiotics.* (vol. 6, pp. 247-291). Springer, Dordrecht. https://doi.org/10.1007/978-94-015-8285-8_12

Kane, D. (2009). *The Kitan language and script*. Leiden/Boston: Brill. https://doi.org/10.1163/ej.9789004168299.i-306

Katayama, Y., Uchida, S., & Sakoe, H. (2009).Stochastic model of stroke order variation. *10th International Conference on Document Analysis and Recognition*, pp. 803-807. IEEE. https://doi.org/10.1109/ICDAR.2009.146

Kawahara, S. (2011). Experimental approaches in theoretical phonology. In M. van Oostendorp (Ed.) *The Blackwell companion to phonology: Phonological interfaces* (vol. 4, pp. 2283-2303). John Wiley & Sons. https://doi.org/10.1002/9781444335262.wbctp0096

Keuleers, E., & Balota, D. A. (2015). Megastudies, crowdsourcing, and large datasets in psycholinguistics: An overview of recent developments. *Quarterly Journal of Experimental Psychology*, *68*(8), 1457-1468. https://doi.org/10.1080/17470218.2015.1051065

Kim, Y., Allassonnière-Tang, M., Miton, H., & Morin, O. (2025). The phonology of letter shapes: Feature economy and informativeness in 43 writing systems. *Journal of Memory and Language, 142*, 104620. https://doi.org/10.1016/j.jml.2025.104620

Kordek, N. (2013). *On some quantitative aspects of the componential structure of Chinese characters*. Poznań: Wydawnictwo Rys.

Kushki, A., Schwellnus, H., Ilyas, F., & Chau, T. (2011). Changes in kinetics and kinematics of handwriting during a prolonged writing task in children with and without dysgraphia. *Research in Developmental Disabilities*, *32*(3), 1058-1064. https://dx.doi.org/10.1016/j.ridd.2011.01.026

Lau, D. K. Y. (2020). Validation of constituent logographemes and radicals in Chinese characters using handwriting data. *Behavior Research Methods, 52*, 305-316. https://doi.org/10.3758/s13428-019-01227-z

Law, S. P., & Leung, M. T. (2000). Structural representations of characters in Chinese writing: Evidence from a case of acquired dysgraphia. *Psychologia, 43*, 67-83.

Lewis, S., & Phillips, C. (2015). Aligning grammatical theories and language processing models. *Journal of Psycholinguistic Research, 44*(1), 27-46. https://doi.org/10.1007/s10936-014-9329-z

Li, Q. L., Bi, H. Y., Wei, T. Q., & Chen, B. G. (2011). Orthographic neighborhood size effect in Chinese character naming: Orthographic and phonological activations. *Acta Psychologica*, *136*(1), 35-41. https://doi.org/10.1016/j.actpsy.2010.09.012

Liu, F., & Wang, Y. (2004). Research on the Khitan large script 'Epitaph for Yelü Changyun'. *Yenching Journal of Chinese Studies, 17*, 61-99.

Meletis, D. (2019). The grapheme as a universal basic unit of writing. *Writing Systems Research, 11*(1), 26-49. https://doi.org/10.1080/17586801.2019.1697412

Miton, H., & Morin, O. (2021). Graphic complexity in writing systems. *Cognition, 214*, 104771. https://doi.org/10.1016/j.cognition.2021.104771

Morin, O. (2018). Spontaneous emergence of legibility in writing systems: The case of orientation anisotropy. *Cognitive Science, 42*(2), 664-677. https://doi.org/10.1111/cogs.12550

Myers, J. (2019). *The grammar of Chinese characters: Productive knowledge of formal patterns in an orthographic system*. London: Routledge. https://doi.org/10.4324/9781315265971

Nakagawa, S., Johnson, P. C., & Schielzeth, H. (2017). The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface*, *14*(134), 20170213. https://doi.org/10.1098/rsif.2017.0213

Neeleman, A., & van de Koot, H. (2010). Theoretical validity and psychological reality of the grammatical code. In M. Everaert, T. Lentz, H. De Mulder, Ø. Nilsen, & A. Zondervan (Eds.), The linguistics enterprise: From knowledge of language to knowledge in linguistics (pp. 183-212). Amsterdam: John Benjamins. https://doi.org/10.1075/la.150.08nee

Ooms, J. (2025). tesseract: Open Source OCR Engine. R package version 5.2.3. https://doi.org/10.32614/CRAN.package.tesseract

Peirce, J. W., Hirst, R. J. & MacAskill, M. R. (2022). *Building experiments in PsychoPy* (2nd ed.) London: Sage.

Peng, X. (2017). Stroke systems in Chinese characters: A systemic functional perspective on simplified regular script. *Semiotica*, *2017*(218), 1-19. https://doi.org/10.1515/sem-2015-0111

Primus, B. (2004). A featural analysis of the Modern Roman Alphabet. *Written Language & Literacy*, *7*(2), 235-274. https://doi.org/10.1075/wll.7.2.06pri

Puranik, C. S., Petscher, Y., & Lonigan, C. J. (2014). Learning to write letters: Examination of student and letter factors. *Journal of Experimental Child Psychology, 128*, 152-170. https://doi.org/10.1016/j.jecp.2014.07.009

R Core Team (2025). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved July 6, 2025, from https://www.R-project.org/

Rastle, K., Lally, C., Davis, M. H., & Taylor, J. S. H. (2021). The dramatic impact of explicit instruction on learning to read in a new writing system. *Psychological Science, 32*(4), 471-484. https://doi.org/10.1177/0956797620968790

Reinken, N. (2023). Handwritten letters and grammatical structures in German: A framework. *Written Language & Literacy*, *26*(2), 155-187. https://doi.org/10.1075/wll.00077.rei

Shu, H., & Anderson, R. C. (1999). Learning to read Chinese: The development of metalinguistic awareness. In J. Wang, A. W. Inhoff, & H.-C. Chen (Eds.), *Reading Chinese script: A cognitive analysis* (pp. 1-18). Mahwah, NJ: Erlbaum. https://doi.org/10.4324/9781410601483-5

Storkel, H. L., & Lee, S. Y. (2011). The independent effects of phonotactic probability and neighbourhood density on lexical acquisition by preschool children. *Language and Cognitive Processes*, *26*(2), 191-211. https://doi.org/10.1080/01690961003787609

Taylor, J. S. H., Plunkett, K., & Nation, K. (2011). The influence of consistency, frequency, and semantics on learning to read: an artificial orthography paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(1), 60-76. https://doi.org/10.1037/a0020126

Tong, X., & McBride-Chang, C. (2010). Developmental models of learning to read Chinese words. *Developmental Psychology, 46*(6), 1662-1676. https://doi.org/10.1037/a0020611

Tong, X., Tong, X., & McBride, C. (2017). Radical sensitivity is the key to understanding Chinese character acquisition in children. *Reading and Writing*, *30*, 1251-1265. https://doi.org/10.1007/s11145-017-9722-8

Topolinski, S., & Strack, F. (2009). The architecture of intuition: Fluency and affect determine intuitive judgments of semantic and visual coherence and judgments of grammaticality in artificial grammar learning. *Journal of Experimental Psychology: General, 138*(1), 39-63. https://doi.org/10.1037/a0014678

Treiman, R., & Kessler, B. (2022). Statistical learning in word reading and spelling across languages and writing systems. *Scientific Studies of Reading, 26*(2), 139-149. https://doi.org/10.1080/10888438.2021.1920951

Treiman, R., Gordon, J., Boada, R., Peterson, R. L., & Pennington, B. F. (2014). Statistical learning, letter reversals, and reading. *Scientific Studies of Reading*, *18*(6), 383-394. https://doi.org/10.1080/10888438.2013.873937

Tsai, C. H. (2006). Frequency and stroke counts of Chinese characters. Retrieved July 6, 2025, from http://technology.chtsai.org/charfreq/

Van Galen, G. P. (1991). Handwriting: Issues for a psychomotor theory. *Human Movement Science, 10*, 165-191. https://doi.org/10.1016/0167-9457(91)90003-G

Van Sommers, P. (1984). *Drawing and cognition: Descriptive and experimental studies of graphic production processes*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511897672

Vidal, Y., Viviani, E., Zoccolan, D., & Crepaldi, D. (2021). A general-purpose mechanism of visual feature association in visual word identification and beyond. *Current Biology, 31*(6), 1261-1267. https://doi.org/10.1016/j.cub.2020.12.017

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language, 40*(3), 374-408. https://doi.org/10.1006/jmla.1998.2618

Vitevitch, M. S., & Luce, P. A. (2005). Increases in phonotactic probability facilitate spoken nonword repetition. *Journal of Memory and Language, 52* (2), 193-204. https://doi.org/10.1016/j.jml.2004.10.003

Vynckier, I. (2025). How OCR works: A close look at optical character recognition. Retrieved July 6, 2025, from https://how-ocr-works.com/index.html

Wan, Y. (2017). Stroke order. In R. Sybesma, W. Behr, Y. Gu, Z. Handel, C.T.J. Huang & J. Myers (Eds.), *Encyclopedia of Chinese language and linguistics* (vol. 4, pp .212-214). Leiden: Brill. https://doi.org/10.1163/2210-7363_ecll_COM_00000399

Wang, J. C. S. (1983). *Toward a generative grammar of Chinese character structure and stroke order*. Ph.D. thesis, University of Wisconsin-Madison.

Wang, R., Huang, S., Zhou, Y., & Cai, Z. G. (2020). Chinese character handwriting: A large-scale behavioral study and a database. *Behavior Research Methods*, *52*, 82-96. https://doi.org/10.3758/s13428-019-01206-4

Wang, X., Yang, J., Shu, H., & Zevin, J. D. (2011). Left fusiform BOLD responses are inversely related to word-likeness in a one-back task. *Neuroimage, 55*(3), 1346-1356. https://doi.org/10.1016/j.neuroimage.2010.12.062

Watt, W. C. (1988). What is the proper characterization of the alphabet? Part 4: Union. *Semiotica, 70*, 199-241. https://doi.org/10.1515/semi.1988.70.3-4.199

West, A. (2025). Babelstone. Retrieved July 6, 2025, from https://www.babelstone.co.uk/

Westphal-Fitch, G., Giustolisi, B., Cecchetto, C., Martin, J. S., & Fitch, W. T. (2018). Artificial grammar learning capabilities in an abstract visual task match requirements for linguistic syntax. *Frontiers in Psychology, 9*, 1210. https://doi.org/10.3389/fpsyg.2018.01210

Wu, J. (2022). Neurocognition in the reading of Chinese characters. *Journal of Chinese Writing Systems, 6*(3), 167-183. https://doi.org/10.1177/25138502211061605

Xin, Y., & Lyu, R. (2022). Experimental measurement and Markov chain modelling of stroke order intuition. *Journal of Physics: Conference Series, 2289*(1), 012015. https://doi.org/10.1088/1742-6596/2289/1/012015

Yang, R., & Wang, W. S. Y. (2018). Categorical perception of Chinese characters by simplified and traditional Chinese readers. *Reading and Writing*, *31*, 1133-1154. https://doi.org/10.1007/s11145-018-9832-y

Yin, J. J. H. (2016). Chinese characters. In S.W. Chan (Ed.), *The Routledge encyclopedia of the Chinese language* (pp. 51-63). Abingdon: Routledge. https://doi.org/10.4324/9781315675541

Zhang, T., & Ke, C. (2018). Research on L2 Chinese character acquisition. In C. Ke (Ed.) *The Routledge handbook of Chinese second language acquisition* (pp. 103-133). Abingdon: Routledge. https://doi.org/10.4324/9781315670706-6