The Effects of Orthographic Depth on Script Acquisition

Cass L. Brown, Quincy A. Rock
Department of Psychology and Human Factors, 1400 Townsend Drive
Houghton, MI 49931 USA

Abstract

Orthographic depth is a term that describes the relationship between a script's appearance and its sound. English is an example of high orthographic depth because of how often letters change their sound (along with its spelling idiosyncrasies). Previous studies have found that languages with a high orthographic depth are harder to read and learn. We sought to show how scripts with shallow orthography are easier to learn because of their structured shape-sound correlations. To do this, we created a minigame for participants to learn characters from one of three scripts: Tengwar (from J.R.R. Tolkien's legendarium), hiragana (from Japanese), and Aurebesh (from Star Wars). We found that Aurebesh, the orthographically deepest of these scripts, was the easiest to learn as measured by reaction time and accuracy. To see if an explanation of shape-sound relationships would make learning Tengwar easier, we conducted another study with the same minigame by manipulating how much background information participants were given. There was no significant difference in the results of that study beyond an increase in reaction time, perhaps due to additional cognitive load. We speculate that the visual distinctiveness of Aurebesh outweighed its orthographic depth and made it the easiest to learn of the three scripts.

Keywords: Tengwar, hiragana, Aurebesh, scripts, linguistics, reading, learning

Introduction

Background

English spelling and pronunciation is peculiar, to say the least. The technical term for this is to say that English has a deep orthography, in which there is a layer of ambiguity and obfuscation between a word's spelling, its pronunciation, and its meaning. An example of a deeper orthography would be kanji, the Japanese logographic script (*logographic* means each character represents an entire word instead of a letter). It's harder to read—and therefore learn—a language with a deep orthography (Ellis et al., 2004; Bar-Kochva & Breznitz, 2014).

Related to orthographic depth is phonological awareness. Phonological awareness is the ability to pair sounds (phonemes) with a character or sequence of characters (graphemes). The more this skill is developed, the easier it is to learn how to read (Liberman & Shankweiler, 1985; Beck & Juel, 2002; Bdeir et al., 2020). Different character pronunciation patterns are acquired through phonological recoding, which is a self-taught mechanism for how certain character patterns change pronunciation, like the difference of the T in "task" versus "motion" versus "fortune" (Share, 1995; Kyte & Johnson, 2006).

Paired together, orthographic depth and phonological awareness explain much of the individual differences in learning to read (Kyte & Johnson, 2006), although there are a multitude of cognitive modules also involved in the reading process. Letters first need to be recognized as distinct from random scribbles, which diverge into separate cognitive pathways only 150 milliseconds after initial exposure (Finkbeiner & Coltheart, 2009). Then, the corresponding sound needs to be identified through phonological awareness, and the difficulty of that awareness is determined by the language's orthographic depth.

Research Focus

The question at play is how the presence and/or awareness of a structure among a new script affects a learner's ability to pick up said script: Tengwar is an unfamiliar script to most and contains less orthographic depth than English, due to a direct shape-sound correlation with its characters. A character in Tengwar is called a *tengwa* (pl. *tengwar*), and by design its anatomy provides clues about what sound it represents (Tolkien, 1955). Since Tengwar has less orthographic depth, it ought to be easier to pick up than scripts that have more ambiguous or arbitrary assignments. However, the *tengwar* look very similar to one another, which may impede learning.

Hiragana, the Japanese syllabic script, is one such ambiguous script because several of its characters look similar but have no sound relation. Aurebesh, the script used in Star Wars visual media, is far more arbitrary in terms of a shape-sound correlation than Tengwar, but may be easier to pick up due to the visual distinctiveness and relative simplicity of each individual character.

A certain amount of phonological recoding must take place with *tengwar* like quessë, which represents the /k/ sound (hard C or K in English). Given how children can learn to read logographic representations of some English words more easily than the actual English word (Rozin et al., 1971; Smith, 1972), Tengwar is a logical extension of this principle: instead of an entire word, its characters represent individual consonant sounds, like in Arabic. Hiragana, by comparison, represents entire syllables. Some phonological recoding must take place in participants so that they can learn this new relationship between character and sound.

Despite this difference in sound representation, we speculate that participants who are informed about the structure of Tengwar (i.e., its orthographic shallowness) will perform better on a Tengwar recognition task by showing higher accuracy, faster reaction times, and more points (which is a composite measure of both accuracy and reaction time). A consistently fast and accurate reaction time would likely mean that participants are more confident in their response, which therefore means that they've successfully associated the *tengwa* stimulus with a phoneme. Visual similarities between the English N and its corresponding *tengwa* (*númen*), along with M and *malta*, may mean that participants learn these quicker. More complicated *tengwar* like *calma* (representing CH as in "chair") or *quessë* (representing C as in "clam" or K as in "kelp") may be harder for participants to learn because they are not exclusive one-to-one substitutions for English letters. We speculate that Aurebesh, in spite of its arbitrary assignment, may prove to be easier to learn for participants than Tengwar because each character in Aurebesh is simple and visually distinct from others. Hiragana likely will fall somewhere between the other two scripts in terms of its difficulty due to some of its similarly shaped yet unrelated characters.

Methods

Participants

Participants were primarily Michigan Technological University students that saw paper flyers or digital announcements, with other groups of potential participants solicited via email. We did not collect personal, demographic, or identifying information for privacy and relevancy reasons. Potential participants who had a familiarity with any of the scripts in the study were told they could not participate.

Materials

The minigame was designed in-house from scratch and published to a web subdomain: the Tengwar minigame link participants were given after completing the informed consent was psy3001.quincyrock.com, while the link participants received for the hiragana/Aurebesh minigame was psy3002.quincyrock.com.

In an effort to "gamify" the task to make the recognition task more appealing, fun, and or engaging to participants, a points system was used. Points were always visible to participants while they were doing the task, as was a question counter to mark their progress (e.g., question 3/75). After completion of the task, a leaderboard of all participants' scores (plus some dummy participants) was shown, along with average reaction time and accuracy. The leaderboard had no research purpose beyond driving engagement.

Stimuli

The shape of a *tengwa* indicates what type of sound it makes. In this recognition task, only four phonological groups were introduced: plosives, which are sounds that involve holding breath and then expelling it rapidly, like the "p" in *explosive*; fricatives, which are sounds that involve breath hissing through a narrow opening (e.g., between the top teeth and bottom lip), like the "f" in *friction*; nasals, which are sounds that involve breath exiting the nose, like the "n" in *nasal*; and voiced/unvoiced sounds, which are the reason for the difference between *thy* and *thigh* or *dug* and *tug* (voiced sounds use the vocal cords, unvoiced ones do not). In Tengwar, a downward stem represents a plosive sound (or an affricate, which is a combination of a plosive and a fricative sound); an upward stem represents a fricative; no stem represents a nasal; and two bows represent a voiced sound, whereas one bow represents an unvoiced sound.

We used 12 unique *tengwar* as a stimulus set, 10 of which were voiced/unvoiced pairs, and of these pairs, 4 were plosives/affricates. The other pair were 2 nasals. The *tengwar* used were *tinco*, *ando*, *parma*, *umbar*, *quessë*, *ungwë* (plosives), *calma*, *anga* (affricates), *formen*, *ampa* (fricatives), *númen*, and *malta* (nasals). The three attention checks used *anto*, *harma*, and *anca* to prevent overlap with the stimulus set.

Tincotéma	*Parmatéma*	*Calmatéma*	*Quessetéma*
 *tinco* /t/ → T	 *parma* /p/ → P	*calma* /tʃ/ → CH[1]	 *quessë* /k/ → C/K[2]
 *ando* /d/ → D	 *umbar* /b/ → B	*anga* /dʒ/ → J[3]	 *ungwë* /g/ → G[4]
 *súlë* /θ/ → TH[5]	*formen* /f/ → F	 *harma* /ʃ/ → SH/TI[6]	 *hwesta* /k/ → QU[7]
 *anto* /ð/ → TH[8]	 *ampa* /v/ → V	 *anca* /ʒ/ → S/SI[9]	 *unquë* /g/ → GH[10]
 *númen* /n/ → N	 *malta* /m/ → M	 *ñoldo* —	 *ñwalmë* /ŋ/ → NG[11]
 *órë* /r/ → R[12]	 *vala* /w/ → W	 *anna* /j/ → Y[13]	 *vilya* —

Fig 1: *A table of the 24 standard* tengwar *in English mode. Those used in the study are bolded and glossed in green text. Footnotes are for those who cannot read IPA phonological notation.*
[1]CH as in "chair"; [2]C as in "cure" and K as in "kin"; [3]J as in "judge"; [4]G as in "gold"

We used 12 unique hiragana for the hiragana variant: chi, ha, he, ho, ku, ne, no, nu, mu, ra, so, and te. Because they do not have names like Tengwar or Aurebesh characters, they were named according to the NATO phonetic alphabet. These names did not matter for the

study beyond naming/calling the image files in the coding process.

ち *alpha* CHI	は *bravo* HA	へ *charlie* HE	ほ *delta* HO
く *echo* KU	ね *foxtrot* NE	の *golf* NO	ぬ *hotel* NU
む *india* MU	ら *juliett* RA	そ *kilo* SO	て *lima* TE
ひ *mike* (Attention check)	ん *november* (Attention check)	き *oscar* (Attention check)	—

Fig 2: *A table of the 12 hiragana used in the study and the 3 attention check characters.*

We used 12 unique Aurebesh characters for the Aurebesh variant: *cresh*, *cherek*, *grek*, *herf*, *jenth*, *nern*, *nen*, *peth*, *senth*, *shen*, *trill*, and *wesk*.

 *cresh* C	 *cherek* CH	 *grek* G	 *herf* H
 jenth J	 *nern* N	 *nen* NG	 *peth* P
 senth S	 *shen* SH	 *trill* T	 *wesk* W
 enth (Attention check)	 *mern* (Attention check)	 *aurek* (Attention check)	—

Fig 3: *A table of the 12 Aurebesh characters used in the study and the 3 attention check characters.*

Scoring Methodology

Each question presented participants with an image of a stimulus (a Tengwar, hiragana, or Aurebesh character depending on their group). Five answers were provided in the format of "[sound] as in [example]" (e.g., "P as in 'paper'"). Each question was worth 100 pts at baseline. For every incorrect answer selected, the respective answer would turn red and play an incorrect sound. Every incorrect answer would also reduce the points gained from the question by 20%. This was a cumulative effect: selecting 2 incorrect answers would reduce the points gained by 40%, and so on. Possible accuracy bounds per question ranged from 0.2 (selecting all incorrect answers before picking the correct one) to 1 (selecting only the correct answer) in steps of 0.2. While this nonbinary accuracy was important for determining points, we used binary accuracy in our resulting figures: binary accuracy was either a 0 or a 1 (any nonbinary accuracy score less than 1 was converted to 0); binary accuracy better shows the effect of learning on a graph.

On selecting a correct answer, the respective answer would turn green and play a correct sound. If the participant clicked the correct answer in under 5 seconds, they would get a point multiplier as determined by $1 + [1 - (t/5000)]$, where t was their reaction time in milliseconds. If it took longer than 5 seconds for the participant to identify the correct answer, the multiplier would be reset to 1 so they would not actually be deducted points for taking longer than 5 seconds.

This organization allowed theoretical point ranges of 20 pts (5+ second reaction time, all incorrect answers selected) to 200 pts (0 second reaction time, only correct answer selected). An experimenter performing the recognition task determined the practical maximum was ~185 pts when the characters were already known. Participants, since they were learning the characters as they did the task, would rarely score this high on a question.

Points gained per question serve as an approximation for learning because it is a composite measure of accuracy and reaction time. If participants learned the characters, they would recognize them faster and select fewer incorrect answers, thereby gaining more points. The accuracy component of the points system discourages spamming all of the answers to try to take advantage of the reaction time bonus.

Design

Before starting the recognition task, participants were shown some instructions. In all script variants (Tengwar, hiragana, and Aurebesh), the instructions explained how the minigame worked and that they would be rewarded for both speed and accuracy.

The experimental Tengwar group was also provided information about how the shape of *tengwa*—namely its stem direction and number of bows—correlated with its sound group. No direct connections between a specific *tengwa* and an English sound were made, so all groups went in blind. The only difference was that the experimental group had access to more information about the underlying structure of Tengwar as opposed to the Tengwar control group. Hiragana and Aurebesh participants only saw information relevant to how the minigame worked and a little blurb about what script they were assigned.

All minigames were between-subject research. It is possible some participants did multiple scripts, but this would not affect their data beyond familiarity with the minigame system.

Procedure

Once they had completed the informed consent and had access to the link, participants could perform the recognition task whenever they wanted as long as they had a stable internet connection as many times as they wanted (although only three participants took advantage of this). Data queries were structured such that only

responses from a participant's first run were exported for analysis.

In all script variants, each character was seen a total of 6 times, organized into blocks of the 12 unique characters. The order of the characters within the blocks was randomized by every block for every session. In Tengwar variants, questions 15, 38, and 70 were attention checks where participants had to select a certain answer to validate their results. In the hiragana and Aurebesh variants, questions 13, 32, and 64 were attention checks. Given the unfamiliarity with this kind of attention check, we still included data from participants who failed the first attention check (but only the first—if they failed any others, their data were excluded). Outliers in reaction time were also excluded, like ones who took more than a minute per question.

After a question was answered correctly, all other answers except the correct one would turn gray and their text was replaced with an em dash (—). For two seconds, the correct answer remained visible so that participants had a window to associate the answer with the character. Two seconds was determined long enough to enable learning while short enough not to drag the session on too long once participants started getting the hang of the task.

Results

Between Scripts

Initially, we looked at the total points for each group to get a general idea of how difficult each script was. However, total points on their own does not demonstrate any learning curves, so we examined more specific measures.

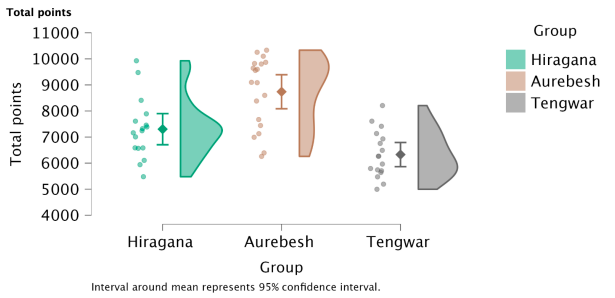


Fig. 4. ***Comparison of means and point distributions across scripts.*** Hiragana: M = 7302.06, 95% CI [6704, 7900]; n = 17. Aurebesh: M = 8738.42, 95% CI [8085, 9391]; n = 19. Tengwar: M = 6326.82, 95% CI [5863, 6790]; n = 17.

We then looked at whether the points gained per question were statistically different between groups. The repeated measures ANOVA showed that points were, in fact, statistically significant between the different script groups.

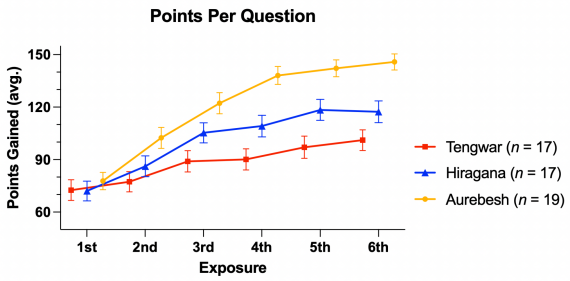


Fig. 5. ***Comparison of points per question across exposures by script.*** A repeated measures ANOVA was performed to compare the effects of script structure on points. There was a statistically significant difference in average points per question between Tengwar, hiragana, and Aurebesh: $F(2,5) = 77.65$, $p < .001$. Error bars are 95% CI.]

Following the points analysis, we looked into both the accuracy and reaction times for each script to see the reason why the points were significantly different from each other. Our data showed that the Aurebesh group was both faster and more accurate on average per question. This is shown by the Aurebesh group's very different mean scores in reaction time and accuracy, along with very small confidence intervals that rarely—if ever—overlap with the other scripts' CIs.

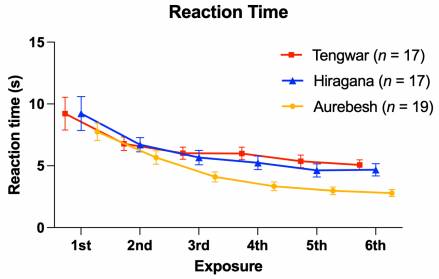


Fig. 6. ***Comparison of reaction times across exposures by script.*** A repeated measures ANOVA was performed to compare the effects of script structure on Reaction time. There was a statistically significant difference in reaction time between Tengwar, hiragana, and Aurebesh: $F(2,5) = 27.43$, $p < .001$. Error bars are 95% CI.

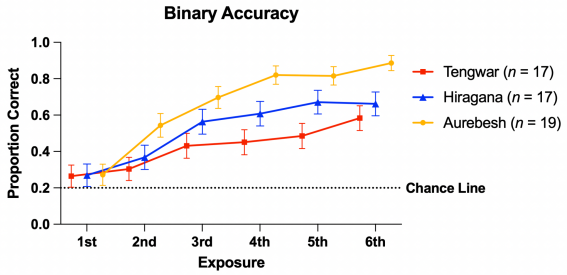


Fig. 7. ***Comparison of binary accuracy across exposures by script.*** A repeated measures ANOVA was performed to compare the effects of script structure on binary accuracy. There was a statistically significant difference in accuracy between Tengwar, hiragana, and Aurebesh: $F(2,5) = 45.42$, $p < .001$. Error bars are 95% CI.

Within Script

We took a similar approach when analyzing the two Tengwar groups (the experimental manipulation being the presence/absence of information regarding Tengwar's structure). Looking at the total points, this analysis was not as immediately useful as in the other study since they were not statistically different.

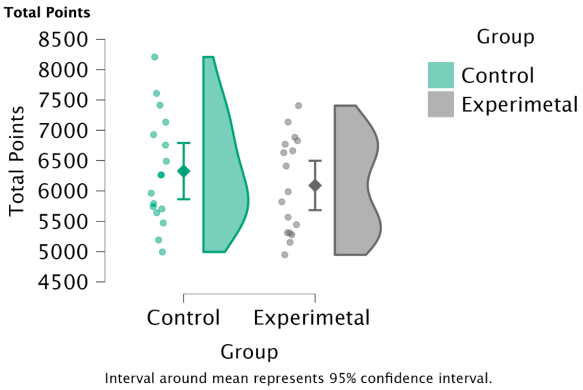


Fig. 8. ***Comparison of means and point distributions within script.*** Control: $M = 6326.82$, 95% CI [5863, 6790]; $n = 17$. Experimental: M = 6090. 95% CI [5683, 6497]

We then looked at the average points per question, which looked very similar to each other as well. This means that either they were the same or each group scored better in one category while tanking in the other (e.g., fast reaction time with poor accuracy or vice versa).

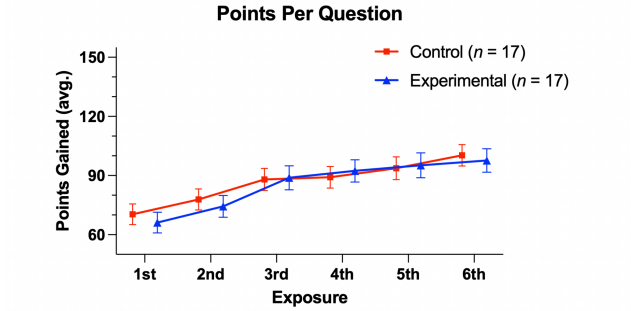


Fig. 9. ***Comparison of points per question across exposures by knowledge groups.*** A repeated measures ANOVA was performed to compare the effect of knowledge of script structure on points. There was not a statistically significant difference in accuracy between the control and experimental groups: $F(1,5) = 1.822$, $p = 0.178$. Error bars are 95% CI.

While looking at the individual accuracy and reaction time, we discovered that the experimental group took significantly longer to answer each question. The reason this wasn't reflected in the points per question despite the similar accuracy levels between groups is that reaction time was only reflected in the points gained if the participant answered in under five seconds, and that was a rare occurrence in this dataset.

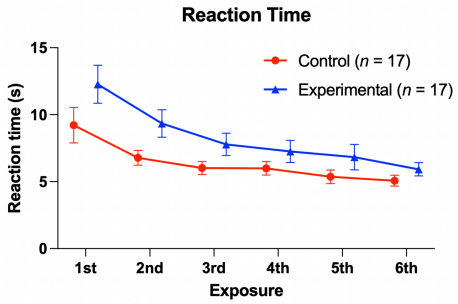


Fig. 10. ***Comparison of reaction times across exposures by knowledge groups.*** A repeated measures ANOVA was performed to compare the effect of knowledge of script structure on reaction time. There was a statistically significant difference in reaction time between the control and experimental groups: $F(1,5) = 22.17$, $p < .001$. Error bars are 95% CI.

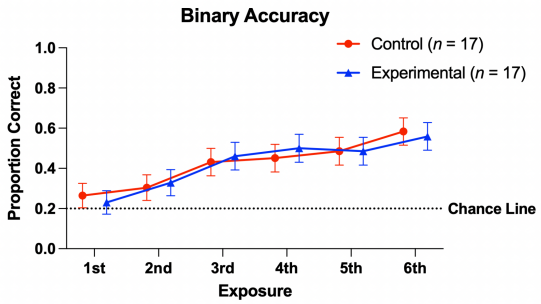


Fig. 11. ***Comparison of binary accuracy across exposures by knowledge groups.*** A repeated measures ANOVA was performed to compare the effect of knowledge of script structure on accuracy. There was not a statistically significant difference in accuracy between the control and experimental groups: $F(1,5) = 0.143$, $p = 0.705$. Error bars are 95% CI.

Discussion

First of all, it is clear from the figures that participants were learning, albeit at different rates depending on the script. Reaction times went down and accuracy went up, which when combined are two clear indicators of increasing recognition. At the very least, this shows that the minigame functioned as intended to facilitate learning.

As seen in **Fig. 1**, *tengwar* are very visually similar to one another because of their orthographic shallowness. This can make them hard to tell apart at a glance, especially when learning them for the first time. We believe this was the primary reason why Tengwar participants showed such little improvement compared to their counterparts in other scripts.

Aurebesh, on the other hand, has very visually distinct characters while still remaining relatively simple, and so we think this is why they were the easiest to learn. Kanji may be visually distinct like Aurebesh characters, but we believe the complexity of the kanji would make them more difficult to learn.

Hiragana were, as expected, a middle ground between the three scripts: the characters are visually distinct to a degree, but some have slight differences that can make identifying the correct sound difficult, and it seems this was reflected in the data.

We noticed that in the binary accuracy figures (**Fig. 5** & **Fig. 8**), all groups scored slightly higher than chance on the first exposure. This is likely because participants learned what answers could NOT be correct for the remaining characters in the first sequence. For example, if you correctly guess "P as in 'peer'" for the Tengwar character *parma*, then you'd know to disregard a potential answer like "P as in 'paw'" for the following characters in the sequence. Knowing one answer is incorrect would make it a 25% chance at guessing, and knowing two incorrect answers would make it a ~33% chance. We believe this is why the first exposure average accuracy was pulled up slightly above chance levels.

As for within-script, the primary difference between the two Tengwar groups' results was their reaction time on the first half of exposures. We speculate that the experimental group took longer on average to respond because they had been given extra information to consider in their instructions. Thinking about how the sound correlated to shape likely slowed them down. However, as can be seen by accuracy and total points per question, this did not significantly affect either of these categories.

It seems likely that the information given about these very similarly shaped characters was not helpful in the context of this task. Instruction in Tengwar would be a setting where this shape-sound information is very helpful, since it allows the characters to become more distinct from one another, but in the environment of a scored minigame, all the information did was create hesitation and/or confusion in the experimental group. Explicit instruction in this special shape-sound correlation rather than a block of text would probably yield better results in the Tengwar recognition task.

The results show that orthographic depth does not play a significant role in the acquisition of a new script *when there is no instruction involved*. Research has only shown this effect of orthographic depth in a classroom setting (Ellis et al., 2004; Bar-Kochva & Breznitz, 2014; Rozin et al., 1971; Smith, 1972). The minigame was designed so that participants had to associate the characters on their own and learn implicitly as they went. It appears that the most visually distinct characters would be the easiest to remember and therefore associate with a sound (i.e., learn), especially when there are 12 unique characters presented in the minigame.

Acknowledgments

The authors would like to acknowledge the free website Tecendil for the ability to transcribe into Tengwar digitally. The authors would also like to thank Dr. Jeffrey Pettibone for his encouragement and advice in this project.

Author Contributions

Brown and Rock jointly thought of ideas and hypotheses for the study design. Rock developed the code for the minigame and drafted most of the manuscript. Brown conducted all data analysis and drafted the results section.

References

Bar-Kochva, I., & Breznitz, Z. (2014). Reading scripts that differ in orthographic transparency: A within-participant-and-language investigation of underlying skills. *Journal of Experimental Child Psychology*, *121*(1), 12–27. doi.org/10.1016/j.jecp.2013.07.011

Bdeir, M., Bahous, R., & Nabhani, M. (2020, June 26). Improving reading readiness in kindergarten children through early phonological awareness interventions. *Education 3-13*, *50*(3), 346–360. doi.org/10.1080/03004279.2020.1851740

Beck, I. L., & Juel, C. (2002). The Role of Decoding in Learning to Read. *The American Educator*. d1wqtxts1xzle7.cloudfront.net/42684524

Ellis, N.C., Natsume, M., Stavropoulou, K., Hoxhallari, L., Van Daal, V.H.P., Polyzoe, N., Tsipa, M.-L., & Petalas, M. (2004). The effects of orthographic depth on learning to read alphabetic, syllabic, and logographic scripts. *Reading Research Quarterly*, *39*(1), 438–468. doi.org/10.1598/RRQ.39.4.5

Finkbeiner, M., & Coltheart, M. (2009). Letter recognition: From perception to representation. *Cognitive Neuropsychology*, *26*(1), 1–6. doi.org/10.1080/02643290902905294

Kyte, S. C., & Johnson, C. J. (2006). The role of phonological recoding in orthographic learning. *Journal of Experimental Child Psychology, 93*(2), 166–185. doi.org/10.1016/j.jecp.2005.09.003

Liberman, I. Y., & Shankweiler, D. (1985). Phonology and the Problems of Learning to Read and Write. *Remedial and Special Education*, *6*(6), 8–17. doi.org/10.1177/074193258500600604

Rozin, P., Poritsky, S., & Sotsky, R. (1971). American children with reading problems can easily learn to read English represented by Chinese characters. *Science*,

171(3977), 1264–1267. doi.org/10.1126/science.171.3977.1264

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

Smith, F. (1972). Phonology and Orthography: Reading and Writing. *Elementary English*, 49(7), 1075–1088. jstor.org/stable/41387875

Tolkien, J. R. R. (1955). Appendix E: Writing and Spelling, § 1, subsec. ii. In *The Return of the King*. George Allen & Unwin.