Crossley, S. A., & Choi, J. S. (in press). Measuring Phonological Complexity Using the Perfect Rhymes Dictionary (PeRDict). *Reading Psychology*.

Title: Measuring Phonological Complexity Using the Perfect Rhymes Dictionary (PeRDict)

Abstract: This paper examines links between perfect rhymes and text readability and decoding using a measure of English rhymes called the Perfect Rhymes Dictionary (PeRDict). PeRDict is based on the Carnegie Mellon University Pronouncing Dictionary (the CMUdict) and provides rhyme counts for ~48,000 words in English and for the most frequent 1,000, 2,500, 5,000, and 10,000 rhymes within the dictionary as measured by the Corpus of American English (COCA). Two assessments of PeRDICT are presented. The first examines the strength of rhyme features to predict text readability in conjunction with word neighborhood density effects reported by the English Lexicon Project (ELP) and a word frequency measure. The second examines the strength of rhyme features to predict word decoding in conjunction with word neighborhood effects and a word frequency measure. In both assessments, the number of rhymes per word was predictive of text or word processing beyond features related to word neighborhood effects and word frequency. Word rhyme counts performed more strongly in predicting text processing versus word processing.

Key Words: reading comprehension, decoding, natural language processing, rhymes, phonological awareness

Phonological awareness is a critical skill for acquiring reading skills in English, with multiple studies demonstrating that early readers that have greater proficiency in recognizing and manipulating syllable structures and individual phonemes are generally more successful readers (Gillon, 2004; Cassady, Smith, & Putman, 2008; Lonigan et al., 2009; Pufpaff, 2009; Schaefer et al., 2009). Much of this is due to links between phonological awareness and word decoding (i.e., correctly pronouncing written words) in English. So many studies have been conducted that demonstrate links between phonological awareness, decoding, and reading ability that some researchers suggest the phenomenon is over-researched (Katz, Brancazio, Irwin, Katz, Magnuson, & Whalen, 2012).

While phonological awareness may be over-researched, there is little research on how the syllable structures of English words related to phonological awareness can be used to predict text readability and word decoding. Perhaps the most common method available for assessing the syllable structure of words in relation to phonological awareness is the quantitative norms for neighborhood density effects reported in the English Lexicon Project (ELP) database (Balota et al., 2007). These norms provide counts for the number of orthographic, phonologic, and phonographic neighbors contained in individual English words. In their simplest form, the counts represent how many words (neighbors) can be formed from an existing word by changing the identity or position of a letter or sound in the word. Studies have indicated that words that come from denser neighborhoods (i.e., words with more neighbors) are processed/decoded more quickly (Andrews, 1989, 1992, 1997; Peereman & Content, 1995).

One problem with measures of neighborhood effects is that they aggregate many features into a single index when indexing the phonological structure of words. For instance, the word *love* will include the rhyme neighbor *shove* and the non-rhyme neighbors *cove, live,* and *lobe*, but it will not include the rhyme *above*. This means that the measures will not fully represent the rhyme-level complexity of words although it is well-established that 1) rhymes are an important component of phonological awareness and help beginning readers realize the syllabic and phonemic structures of language (Gunning, 2001; Konza, 2011), and 2) rhymes as neighbors are common phenomena, especially for monosyllabic words (de Cara & Goswami, 2002).

The purpose of this study is to introduce and test a new method for calculating rhyme counts using the Perfect Rhymes Dictionary (PeRDict)[i]. PeRDict was developed to assess text readability and word decoding and is based on the Carnegie Mellon University Pronouncing Dictionary (the CMUdict). PeRDICT provides rhyme counts for ~48,000 words in English and for the most frequent 1,000, 2,500, 5,000, and 10,000 rhymes within the dictionary as measured by the Corpus of Contemporary American English (COCA; Davies, 2009). This paper discusses the development of PeRDict and presents use-case assessments of the database. The first assessment examines the strength of PeRDict features to predict text readability in a corpus of texts leveled for primary and secondary educational settings. The second assessment examines the strength of PeRDict features to predict word decoding (as measured by lexical decision times) in adults. Both studies assess PeRDict features in conjunction with control variables known to influence text readability and decoding: word neighborhood effects and word frequency.

Phonological Awareness

Phonological awareness is the understanding and ability to use and reflect on the sound structures in a language to assist in the processing of that language (Stuart, 2005, Wagner, 1988) and is an important part of lexical restructuring (Metsala & Walley, 1998). Phonological awareness is particularly important in languages that do not contain a one-to-one mapping between graphemes and phonemes (e.g., English). Previous studies have shown strong links between children's phonological awareness and their ability to decode words and their vocabulary growth (Carroll, Snowling, Hulme, & Stevenson, 2003; Hipfner-Boucher et al., 2014; Lonigan, 2007; McDowell, Lonigan, & Goldstein, 2007). It is theorized that phonological representation for words becomes more specified in children's lexicon to reduce confusion between similar-sounding words as the number of vocabulary words known increases (Elbro & Jensen, 2005; Metsala, 1999). Thus, as children's vocabulary grows, children need to distinguish between more words and begin to rely on syllable and phoneme structures as one strategy to do so (Metsala & Walley, 1998; Walley, 1993).

The phonological structure of words and the effect of that structure on word decoding play an important role in the development of reading skills (Perfetti & Stafura, 2014; Verhoeven, van Leeuwe, & Vermeer 2011) such that phonological awareness is a strong predictor of reading ability (Duncan et al., 1997; Hulme, 2002; Melby-Lervåg, Lyster, & Hulme, 2012; Nithart et al., 2011; Pfost, 2015; Schatschneider, Fletcher, Francis, Carlson & Foorman, 2004; Stuart, 2005; Swanson & Alexander, 1997). Researchers have argued that phonological awareness in a language like English assists early readers in learning how to read by helping them to discern which letters correspond to the onsets and rhymes found in spoken language (Goswami & Bryant, 1990). Research has also indicated that children learning English, who are better at manipulating onsets and rhymes, develop reading proficiency more quickly (Bradley & Bryant, 1983).

Early empirical research linking larger sound units (i.e., syllables), specifically rhyming and alliteration, to the development of reading skills in English reported strong associations between measures of phonological awareness and reading skills (Bradley & Bryant, 1978;

Calfee, Lindamood, & Lindamood, 1973). Similar research into rhymes indicated that stronger phonological awareness in young children was predictive of reading achievement in later years (Blachman, 1984; Bradley, & Bryant, 1983; Fox & Routh, 1975; Share, Jorm, MacLean, & Matthews, 1984) and that phonological awareness likely aided in the acquisition of both spelling and reading skills (Ball, & Blachman, 1991; Bradley & Bryant, 1983; Lewkowicz, 1980; Lundberg, Frost, & Peterson, 1988; Bryant, Maclean, Bradley, & Crossland, 1990).[ii] However, there is no agreement on which features of phonological awareness are the most important for understanding the reading process. There is also no real agreement about how best to operationalize features related to phonological awareness (Anthony et al., 2002; Morais, 2003; Stanovich, Cunningham, & Cramer, 1984; Yopp, 1988).

Measuring Phonological Awareness

Knowledge of syllable structure can be assessed in children through behavioral tasks designed to detect rhymes or generate words that rhyme. For instance, behavioral assessments may ask young children to recognize sequences of single sounds by deleting and replacing onset syllables and phonemes to create rhymes (Gunning, 2001; Konza, 2011) or to identify which words rhyme or produce rhymes (Muter et al., 1998). Non-rhyming tasks that assess knowledge of syllable structure ask children to delete a syllable from a word that is not part of the onset and produce the remaining sequence of sounds (Catts et al., 2001). Assessing phonemic awareness may also include asking children to isolate, blend, segment, and manipulate phonemes (Konza, 2011; Stanovich, Cunningham, & Cramer, 1984).

Another approach to assessing phonological awareness is through neighborhood density effects. Neighborhood density effects examine how many words can be created from a single word by changing one character (in the case of orthographic neighbors), one phoneme (in the case of phonological neighbors), and one letter and phoneme where the correspondence between the two is exact (in the case of phonographic neighbors). Neighborhood effects lend themselves to quantitative assessments because word neighbors can be calculated based on existing dictionaries.

Early work by de Cara and Goswami (2002) developed a dataset for calculated neighborhood density effect for ~4,000 monosyllabic words found in the CELEX corpus (Baayen, Piepenbrock, & Gulikers, 1993). They reported that most phonological neighbors for words that had high density (i.e., words with many neighbors) were rhymes (~60%), but that this number trailed off for words with lower density (~42%). A smaller sample of monosyllabic words (n = 632) with rhymes were compared to their age of acquisition scores reported by Gilhooly and Logie (1980). De Cara and Goswami (2002) found that the percent of rhyme neighbors increased as age of acquisition scores increased.

The English Lexicon Project (ELP) database (Balota et al., 2007) also contains neighborhood density counts for phonological, orthographic, and phonographic neighbors, but these are not separated for rhymes. Studies using the ELP norms have found that words that come from denser orthographic neighborhoods (i.e., words with a lot of neighbors) are processed/decoded more quickly (Andrews, 1997; Kim, Crossley, & Skalicky, 2018)[iii] and this effect is stronger for low frequency words than high frequency words (Andrews, 1989, 1992; Peereman & Content, 1995). Similar findings have been reported for words with denser phonological neighborhoods (Kyle, Crossley, & Berger, 2018; Skalicky, Crossley, & Berger, 2019; Yarkoni et al., 2008).

Current Study

There is no strong agreement about which features of phonological awareness are the most important for understanding reading or decoding processes. There is also little agreement about how to operationalize features that measure phonological awareness (Anthony et al., 2002; Morais, 2003; Stanovich, Cunningham, & Cramer, 1984; Yopp, 1988). Lastly, few quantitative features beyond the neighborhood density effects reported by de Cara and Goswami (2002)[iv] or found in the ELP have been developed to assess phonological awareness and none have been developed that directly assess word rhymes on a large scale.

To address these concerns, we introduce the Perfect Rhymes Dictionary Database (PeRDict), which is based on the CMU Pronouncing Dictionary (CMUDict). The PeRDICT database comprises around 48,000 words and contains metrics for the number of rhymes for each word and for rhymes based on word frequency. The goal of the PeRDICT database is to provide a metric to measure the number of rhymes associated with individual words. A database of rhymes for the English language allows for the examination of associations between quantitative measures that tap into phonological awareness and reading/writing skills in learners, the investigation of rhyme development over time in learners, and the potential to control for potential rhyming effects in behavioral stimuli used in psycholinguistic and other types of reading experiments. We assess the PeRDICT database through two research questions that guide the study.

1. Can the rhymes reported in the PeRDICT database predict text readability in a K-12 corpus in combination with neighborhood effects and word frequency measures ?
2. Can the rhymes reported in the PeRDICT database predict word decoding for adults as measured in lexical decision task in combination with neighborhood effects and word frequency measures ?

We hypothesize that rhymes will play an important and unique role in predicting both text readability and word decoding. However, since both of these tasks rely on lexical processing, we presume that word frequency counts that strongly measure exposure to language will be stronger predictors (Balota & Chumbley, 1984; Rayner & Duffy, 1986). We also presume that since reading and decoding tasks rely on the written word, neighborhood density effects based on orthography will be stronger predictors.

Methods

Our methods include the selection of control variables to include in the statistical analysis. The control variables are meant as moderators to ensure that features found in the PeRDICT database are predictive of readability when features related to word frequency and neighborhood effects are included in predictive models. The methods also include the development of the PeRDICT database and the selection of comparison data that includes a readability corpus and a word decoding database. Lastly, the methods include the statistical analyses used in assess the predictive strength of the PeRDICT database and the control variables.

Control Variables

We first discuss the control variables used when assessing the PerDICT database. The control variables selected (word frequency and word neighborhood measures) are commonly used to assess word decoding in reading tasks.

Word Frequency. We include a measure of word frequency in our analyses of phonological awareness counts as a co-variate because word frequency is a strong predictor of both text readability and word decoding (Balota & Chumbley, 1984; Rayner & Duffy, 1986). We derived our frequency norm from SUBTLEXus, which is a corpus of subtitles from 8,388 American films and television shows that comprise 51 million words (Brysbaert & New, 2009).

We specifically selected word frequency scores from SUBTLEXus that were logarithmically transformed (log10) to control for Zipfian distributions common to frequency measures (Zipf, 1935).

Word Neighborhood Measures. We derived word neighborhood measures from the English Lexicon Project (ELP) database. The ELP database provides descriptive and behavioral data for around 80,000 English words, including proper names and contractions. The database provides numeric representations for words with reference to orthographic, phonologic, and phonographic neighbors (see below for more details), word frequency metrics based on Hyperspace Analogue to Language (HAL, Lund & Burgess, 1996) corpus, and word naming and lexical decision times (see below for more details). For each of these features, the percentage of words represented differs. For instance, all words have representative counts for some neighborhood counts, while counts for frequency are available for between ~22,000 and ~38,000 words and word naming and lexical decision times are available for ~40,000 words.

*Orthographic neighbors***.** The ELP database contains measures related to orthographic neighborhood effects for individual words (excluding homonyms). The simplest measure is based on Coltheart's N (i.e., Ortho_N; Coltheart, Davelaar, Jonasson, & Besner, 1977), which is the number of alternate words that can be generated by changing a single letter in a word. Ortho_N is a calculation of the number of words that can be created by changing one letter in a word without changing the identity or the positions of the other letters in the word. For example, the word *cat* includes the orthographic neighbors *oat, cot, vat, cab, chat, mat, cam, bat, rat, cad, hat, cap, pat, fat, flat, frat, sat, eat, car, cut*, and *can* among others (Balota et al., 2007). Notice that while many of the neighbors are rhymes of *cat* (e.g., *vat, mat, rat*), the majority are not.

Additionally, it is important to note that neighborhood effects do not capture all rhymes. For instance, the rhyme *combat* would not be counted as a neighbor.

The ELP database also reports the mean frequency of the orthographic neighbors per individual words (Freq_O). So, for the example above, the log frequency for all the neighbors of *cat* is computed using word frequencies derived from HAL corpus. The ELP database also includes orthographic similarity effects based on Levenshtein distance (Levenshtein, 1966) per word. The measure, called orthographic Levenshtein distance 20 (OLD20, Yarkoni, Balota, & Yap, 2008) measures the minimum number of substitution, insertion, or deletion operations needed to change one word into another and calculates the mean LD for a word in relation to its 20 closest neighbors. This measure addresses two limitations of Ortho_N. First, Ortho_N is a binary measure such that two words are either neighbors or not regardless of perceptual similarity between words. Second, Ortho_N relies on single letter substitution and does not allow for insertions, deletions, or transposition operations. Using OLD20 calculation, but not Ortho_N calculations, words like *widow* and *window* and *trail* and *trial* are orthographically similar.

Phonological neighbors. The ELP database also contains phonological neighborhood features that exclude homonyms. Unlike orthographic neighbors, phonological neighbors are based on phonemes. For instance, *sh* is recognized as a single phoneme, whereas orthographically, it is two letters (s and h). So, *shove* and *stove* would not be related phonologically, but they would be orthographic neighbors. The ELP calculates the same features for phonological neighbors as it does for orthographic neighbors. These include measures based on Coltheart's N (i.e., Phono_N), the mean frequency of the phonological neighbors per individual word (Freq_P), and phonological neighbors based on Levenshtein distance (PLD).

Phonographic neighbors. Phonographic neighborhood effects in the ELP database measure lexical print-sound conversions excluding homophones. Phonographic neighbors are based on words that differ by one letter and one phoneme. Thus, *stove* can be converted to *stone* because *v* and *n* are both letters and phonemes. The difference between *stove* and *shove*, however, is not a phonographic conversion because *s*, *t*, and *sh* are phonemes. However, this conversion would be considered an orthographic neighbor because *t* can be replaced with *h* as letters. Like orthographic neighbors, the features for phonographic neighbors in ELP include a measure based on Coltheart's N (i.e., PG_N). ELP also reports the mean frequency of the phonographic neighbors per words using word frequencies derived from the HAL corpus (Freq_PG). ELP does not include phonographic neighbors based on Levenshtein distance.

The Perfect Rhymes Dictionary (PeRDict) Database

We used the CMU Pronouncing Dictionary (the CMUdict) to develop the Perfect Rhymes Dictionary (PeRDict). The CMUdict is an open-source pronouncing dictionary developed at Carnegie Mellon University to assess rhymes. The CMUdict provides a mapping between a word and that word's phonemes for 133,779 words in the English language. However, in practice, many of these words are extremely rare. Because of the rarity of many words in the CMUdict, we elected to only keep those words that overlap between the CMUdict and the English Lexicon Project (ELP). The ELP comprises 79,672 words, of which 47,885 words were shared with the CMUdict. These ~48,000 words were selected for initial inclusion into PeRDict.

The CMUdict includes the word (the key) and pronunciations in the form of phonemes (the value). Stress marks (as numbers) are also included for stressed vowels in each word wherein vowels followed by the number 1 are stressed. An example of the CMUdict for four words (*nap, map, maps, collapse*) is provided in Table 1.

[Insert Table 1 here]

To calculate perfect rhymes for the ~48,000 selected words, we considered words to be perfect rhymes of other words when the stressed vowel in both words in the CMUdict was identical as well as any following sounds (i.e., the rhyme). Additionally, the consonant(s) before the stressed vowel (if present) had to differ (i.e., the onset was different). Thus, *nap* and *map* in Table 1 would rhyme as would *maps* and *collapse*, but *nap* would not rhyme with *collapse*.

To calculate the number of perfect rhymes for the ~48,000 words shared between the CMUdict and the ELP database, we wrote a Python script that measured the number of words that shared the same perfect rhymes with each word. Thus, each word was compared to the other ~48,000 words and raw counts were computed for the number of words that rhymed with each word in the CMUdict.

As noted above, many of the ~48,000 words shared between the CMUdict and the ELP database are rare and thus not likely candidates to influence word decoding (i.e., children are unlikely to have exposure to these rare words, so they probabilistically would not influence the development of rhymes and the decoding of words). To control for the rarity of words and the effect this could have on rhyme development and word processing, we identified the most frequent content words in the English language using the Corpus of Contemporary English (COCA). We ignored function words because their irregular spellings and reduced stress likely affect their rhyme patterns and learnability (Weber, 2006). COCA contains more than one billion words from texts produced from 1990-2019 and includes eight genres: subtitles, spoken, fiction, magazine, web pages, blogs, newspapers, and academic texts. From COCA, we selected the most frequent content words for the following frequency bands: 1,000 words, 2,500 words, 5,000 words, and 10,000 words. We removed words that were not in these bands to control for rare words in the CMUdict when calculating the number of rhymes. Removing rare words helped ensure that the PeRDICT database included rhyme counts based on words that are likely known by readers and would more likely influence word decoding.

To calculate the number of perfect rhymes for the reduced word lists extracted from COCA based on frequency (i.e., the 1,000, 2,500, 5,000, and 10,000 most frequent words), we followed a similar approach used with all words shared between the CMUdict and the ELP database. For example, for the 1,000-word list, each word was compared to the other ~1,000 words and the number words that rhymed for each word within that 1,000-word list was computed.

For all rhyme counts, we calculated the number of rhymes for the raw words (i.e., tokens) and for word lemmas. Word lemmas differ from tokens because the lemmas remove morphological elements of the word to better represent the basic form of the word. Using our example from before, the words *maps* and *collapse* would rhyme at the token level, but they would not rhyme with *nap*. However, during lemmatization, *maps* would revert to *map* and would then rhyme with *nap*. To lemmatize the words, we used the natural language processing package spaCy (Honnibal et al., 2020).

Not all words in the CMUdict had stressed vowels (e.g., *y'all, marketers, greedier* and *priciest*), which resulted in 20 words shared between the CMUdict and the ELP database being removed. Thus, the final count for words shared between the CMUdict and the ELP database for which all perfect rhymes could be calculated was 47,865. For the COCA word frequency lists, there was also not perfect overlap with either the CMUdict or the ELP database, meaning that counts for the most frequent words were not perfectly aligned with the numbers 1,000, 2,500, 5,000, and 10,000 and so our frequency bands are adjusted (see Table 2 for the actual number of words in the perfect rhyme counts for these frequency bands). Also, there were many words that reported zero counts for rhymes for the entire dataset and for the frequency bands. This is because some words in each sub-dataset did not rhyme with any of the other words in that sub-dataset. A good example is the word *travel* which had zero rhymes within the COCA 1,000, 2,500, and 5,000 frequency bands but reported two rhymes in the 10,000 frequency band and four in the total data set. An example of a word that had no rhymes in the entire dataset is *dubious*. The number of zero counts for each dataset is reported in Table 2.

In total, PeRDict includes 12 variables: perfect rhymes calculated for the 47,865 shared words and lemmas between the CMUdict and the ELP database and perfect rhymes calculated for the 994 (1,000-word band), 2,471 (2,500-word band), 4,895 (5,000-word band), and 9,538 (10,000-word band) most frequent words and lemmas reported by COCA.

[Insert Table 2 here]

Comparison Data

To assess the predictive strength of the PeRDICT database, we examined links between the features reported in the database, the orthographic, phonologic, and phonographic features reported in the ELP database, the SUBTLEXus frequency count, and two outcome variables related to text readability and word decoding. The outcome variables are discussed below.

Text readability. To assess the links between the rhyme, orthographic, phonologic, phonographic, and frequency variables and text readability, we used the CommonLit Ease of Readability (CLEAR) corpus (Crossley, Heintz, Choi, Batchelor, & Karimi, 2021; Crossley, Heintz, Choi, Batchelor, Karimi, & Malatinszky, 2023). The CLEAR corpus contains 4,724 reading samples appropriate for primary and secondary students. The corpus was developed to model and test various readability metrics. To collect unique readability scores for each excerpt, kindergarten through 12th grade teachers were recruited via email from CommonLit's internal teacher pool. Teachers were shown two text excerpts at a time and instructed to select which excerpt they believed was easier to read for their K-12 students. After removing outliers, data was kept from 1,116 teachers, who made 111,347 overall comparison judgments. A Bradley-Terry model (Bradley & Terry, 1952) was used to compute pairwise comparison scores for the teachers' judgments of text ease to calculate unique readability scores for each excerpt. The final scores reflect the "Easiness" for each excerpt in the corpus and are labeled BT ease. The CLEAR corpus was separated into a training set (n = 2,834) and a test set (n = 1,890). A Python script was developed to calculate the mean number of rhymes per types and tokens for each excerpt in the CLEAR corpus (normalized by text length).[v]

Word Decoding Data. To assess links between the rhyme, orthographic, phonologic, phonographic, and frequency variables and visual text decoding, we used the lexical decision times reported in the ELP database. Previous research on lexical decision times supports the notion that lexical decision times target decoding-like processes used in printed word identification (e.g., Katz et al., 2012; Lukatela & Turvey, 2000; Rastle & Coltheart, 2006). Unlike the CLEAR corpus above, which taps into K-12 reading skills, the ELP database is derived from adult participants. While rhyming patterns develop in early childhood, it is likely that patterns learned leave a preference for onset-rhyme organization in processing tasks such as decoding (Goswami & Bryant, 1990), which can be assessed in adult data.

The lexical decision times in the ELP database are based on a list of ~40,000 words taken from the Kucera and Francis (1967) and CELEX (Baayen et al., 1993) norms. The list of words included proper names and contractions. Lexical decision times were derived from ~1,300 native English speakers recruited from research participant pools at six universities. Participants saw several blocks of words (usually 250 words per block). In each block, participants would see a string of letters and had to decide if the string of letters was a word in English or not by pressing a button. After responding, they would see another string of letters. Breaks were given in between blocks. Response times for each word based on button pushes were averaged across all participants to provide a single measure of lexical decision time for each of the ~40,000 words.

For our measure of lexical decision time, we selected the z-scored lexical decision variable. For this variable, each participant's raw lexical decision latencies were transformed using z-scores and the mean z-score for all participants per word was calculated, allowing for decision time performance to be compared across words. According to Balota et al. (2007), this is the most reliable metric of lexical decision time because it minimizes processing speed and variability reported among participants.

Of the 47,865 words that were shared between the ELP data and the PeRDict database, only 34,407 had lexical decision time data. This was our starting point for pruning down variables based on word overlap between dictionaries. The ELP variables calculating the number of phonographic neighbors and those calculating orthographic, phonological, and phonographic neighborhood frequency reported fewer than 20,000 words (Min = 11,315, Max = 18,467) that overlapped with words in the PeRDict database. Because of low overlap, these variables were removed. The variables that calculated Levenshtein distances for closest neighbors were available for 33,558 of the words shared between the PeRDict and lexical decision variables, further winnowing the data. Lastly, only 32,105 words were shared between the SUBTLEXus frequency variable and the remaining PeRDict and ELP variables. Thus, our final data frame for the lexical decision time analysis included 32,105 words. Also, because we were only matching words between the PeRDict data frame and the ELP data frame, which includes both lemmas and tokens, no distinction was made between lemmas and tokens as was done in the CLEAR corpus analysis.

Statistical Analyses

To predict text reading ease scores found in the CLEAR corpus and the lexical decision times found in the ELP database, we used the rhyme variables from PeRDICT, the phonological, orthographic, phonographic features reported in the ELP database, and the SUBTLEXus frequency counts as predictor variables in linear models. We ran three linear models.

The first was to predict the ease of reading scores found in the CLEAR corpus. For this analysis, feature counts for the ELP neighborhood variables were calculated using the Tool for the Automatic Analysis of Lexical Sophistication (TAALES, Kyle et al., 2018). TAALES calculates neighborhood counts for the words in each excerpt that had a matching word in the ELP database and then normed that count by the number of words in the text.

For the lexical decision times, we ran two linear models. The first examined the PeRDict rhyme counts for 32,105 words shared between the SUBTLEXus, ELP, and PeRDict variables. The second linear model examined the adjusted 10,000 most frequent words shared between PeRDict and the ELP corpus. In practice, this analysis examined the 8,464 words shared between the COCA 10,000 most frequent words, the CMUDict that informed PeRDict, and the 32,105 words extracted from the ELP and the SUBTLEXus databases. This analysis was conducted to ensure that all rhyme variables reported a value (i.e., the rhyme counts for the most frequent 1,000 words would only report a value for ~1,000 words and not all 32,105 words). Additionally, previous studies have indicated that neighborhood frequency effects are stronger for low frequency words compared to high frequency words (Andrews, 1989, 1992; Peereman &

Content, 1995) and running two regression analyses could tap into these differences, if they exist.

In each linear model, we first calculated bivariate Pearson correlations using the cor.test() function in R (R Core Team, 2020) to identify highly collinear features among the rhyme variables and the phonological, orthographic, phonographic and frequency variables for both the CLEAR reading ease scores and the lexical decision times. If two or more variables correlated at $r > .699$, the rhyme, phonological, orthographic, phonographic and/or frequency variable(s) with the lowest correlation with the ease of readability score or the lexical decision time was removed and the variable with the higher correlation was retained. The threshold ($r > .699$) was chosen because it is a widely used metric that research shows to be an indicator for when collinearity may severely affect a models' integrity (Dormann et al., 2013; Crossley & Kyle, 2022). We also only retained variables that demonstrated at least a small relationship with the ease of readability scores or the lexical decision times ($r > .099$).

We used the CARET package (Kuhn, 2008) in R to develop linear models for both the reading ease scores and the lexical decision times. Model training and evaluation were performed using a training set and a test set in which the training set comprised 70% of the text excerpts or lexical decision times randomly selected. Estimates of accuracy are reported using the amount of variance explained by the developed models (R^2). The models were hand-pruned to remove non-significant variables. The relative importance of the indices in each model was calculated using the calc.relimp() function in the relaimpo package (Grömping, 2006) using the lmg metric (Lindeman, Merenda, & Gold, 1980). lmg takes into account both the direct relationship between the independent and dependent variable (i.e., the bivariate correlation) and the indirect relationship between the independent and dependent variable (i.e., the amount of variance explained when included in a multivariate model).[vi]

Results

Our statistical analyses comprise two sections. The first examines the strength of the PeRDICT database features along with variables related to neighborhood effects and word frequency to predict text readability in the CLEAR corpus. To do this, we conduct simple correlations across the variables followed by regression modeling. The second and third analyses examine lexical decision times as a proxy for decoding. These analyses are separated by word frequency with the first comprising the 10,000 most frequent words in the ELP database and the second comprising all words in the ELP database. Like the CLEAR corpus analysis, both the second and third analyses examine the strength of the PeRDICT database features along with variables related to neighborhood effects and word frequency to predict lexical decision times.

CLEAR Corpus Reading Ease Score

Correlations. Of the 25 variables assessed, 19 variables were removed because of multi-collinearity or because they did not report at least a small effect size. Correlations among the remaining variables indicated a strong relationship ($r > .50$) between the frequency variable and reading ease scores. Low to medium relationships ($r > .10$ and $< .50$) were reported for the rhyme and neighborhood density variables with the strongest correlation reported for rhymes in the most frequent 1,000 lemmas in English. Correlations are reported in Table 3.

[Insert Table 3 here]

Linear model. The six variables that remained after controlling for multicollinearity and that showed at least a weak relationship with reading ease scores were entered into a linear model. The best model after hand pruning the variables to remove suppression effects and non-significant predictors included two variables (SUBTLEXus frequency and Rhymes lemma 1,000) and reported r = .539, R^2 = .291, F (2, 3303) = 677.4, p < .001 (see model parameters summarized in Table 4). When the model was applied to the test set it reported r = .560, R^2 = .314. The relative importance metrics indicate that the strongest predictors of reading ease was SUBTLEXus frequency (.530) followed by Rhymes lemma 1,000 (.470). The variables indicated that text readability was predicted by more frequent words and frequent words with a greater number of frequent rhymes. Post-hoc tests indicated all linear model assumptions were met.

[Insert Table 4 here]

Lexical Decision Times for Adjusted 10,000 most Frequent Words

Correlations. Of the 14 variables assessed, eight variables were removed because of multi-collinearity or because they did not report at least a small effect size with lexical decision time. Correlations indicated strong to medium relationships (r > .30) between the remaining phonological and orthographic features and lexical decision times. The rhyme variables all reported small effect sizes (r > .10 and < .30). The word frequency measure was removed from the analysis because it strongly correlated with the Levenshtein distance for phonological neighbors variables but reported a lower correlation with lexical decision time. Correlations are reported in Table 5.

[Insert Table 5 here]

Linear model. The six variables that remained after controlling for multicollinearity and that showed at least a weak relationship with lexical decision scores were entered into a linear model. The best model after hand pruning the variables to control for suppression effects and to remove non-significant predictors included three variables related to rhymes and Levenshtein distance for phonologic neighbors. The model reported r = .567, R^2 = .332, F (3, 5920) = 981.1, p < .001

(see model parameters summarized in Table 6). When the model was applied to the test set, it reported r = .559, $R^2 = .313$. The relative importance metrics indicate that the strongest predictors of lexical decision time was Levenshtein distance for phonological neighbors (.848) followed by Rhymes token 2,500 (.091) and Rhymes token 1,000 (.062). The variables indicated that lexical decision times were best predicted by words with a greater number of phonological neighbors and words with more rhymes. Post-hoc tests indicated all linear model assumptions were met.

[Insert Table 6 here]

Lexical Decision Times All Words

Correlations. Of the 14 variables assessed, six variables were removed because of multi-collinearity or because they did not demonstrate at least a small effect size. Correlations among the non-multicollinear variables indicated a strong relationship ($r > .50$) between the frequency variable and lexical decision times. The neighborhood density variables reported strong to medium effects (r > .30) while the rhyme variables demonstrated small effects (r < .20). Correlations are reported in Table 7.

[Insert Table 7 here]

Linear model. The eight variables that remained were entered into a linear model. The best model after hand pruning the variables included three variables related to rhymes, phonographic neighbor density, and word frequency. The model reported r = .732, $R^2 = .536$, $F (3, 22469) = 8656.0$, $p < .001$ (see model parameters summarized in Table 8). When the model was applied to the test set it reported r = .726, $R^2 = .527$. The relative importance metrics indicate that the strongest predictors of lexical decision time was word frequency (.563) followed by phonographic neighborhood density (.410) and the perfect rhymes (.027). The variables indicated that lexical decision times were best predicted by frequent words, words with denser phonographic neighborhoods, and words with more rhymes. Post-hoc tests indicated all linear model assumptions were met.

[Insert Table 8 here]

Discussion

This paper examines links between perfect rhymes and text readability and decoding. It does so using the Perfect Rhymes Dictionary (PeRDict) database, which is based on the CMU Pronouncing Dictionary (the CMUdict). The PeRDICT database provides the number of perfect rhymes for ~48,000 English words as well as the number of perfect rhymes for more frequent words for an adjusted 1,000, 2,500, 5,000, and 10,000 most frequent words as calculated by the Corpus of Contemporary American English. PerDICT was first used to assess links between the number of rhymes a word has and text readability. PerDICT was then used to examine the strength of word rhymes in predicting lexical decision times in all words and the adjusted 10,000 most frequent words. In all cases, the number of rhymes per word was predictive of both text readability and word decoding when including features related to word frequency and/or word neighbors. Word rhyme counts performed more strongly in predicting text processing versus word decoding. In all analyses, a word frequency count or a word neighbor count outperformed rhyme counts.

In terms of predicting text readability, a clear trend indicated that rhyme counts based on more frequent words showed stronger correlations with reading ease scores. For instance, rhymes for the most frequent adjusted 1,000 lemmas showed a correlation of .448, and this steadily decreased at each frequency level. Similar trends were reported for rhymes based on tokens. When rhyme count features were used in a regression model to predict readability scores for the text excerpts, along with neighborhood effect and word frequency features, two features related to word frequency and rhymes for the adjusted most frequent 1,000 lemmas explained 31% of the overall variance in the test set. The word frequency measure explained 53% of the variance, while the rhyme variable explained 47% of the variance. Excerpts that were easier to read had words that were more frequent and included more frequent rhymes. Thus, a strong link between word rhymes and the readability of a text was found, one that explained a significant amount of variation beyond a measure of word frequency. Of note, no measures of phonological, phonographic, or orthographic neighborhood densities were significant predictors of readability in the regression model, although they did demonstrate strong correlations.

In terms of predicting individual word decoding, a weaker trend in the rhyme counts was reported when compared to predicting text readability. In both linear regressions (the adjusted most frequent 10,000 words and all words), most rhyme counts for tokens (absent the rhyme counts for the full PeRDict) showed at least a small effect with word decoding with the strongest correlation reported for rhyme counts for words with the 2,500 and 5,000 most frequent words in English respectively (at -.254 for the adjusted 10,000 most frequent words and -.195 for all words). When rhyme count features were used in a regression model to predict word decoding times for the adjusted 10,000 most frequent words, along with neighborhood density features, one feature related to Levenshtein phonological neighbor distance and two rhyme features for the most frequent 1,000 and 2,500 words explained 33% of the overall variance. The Levenshtein distance for phonological neighbors explained ~85% of the variance while the rhyme count variables explained the remaining 15%. The finding indicated that words with a greater number of neighbors derived from inserting, deleting, or transposing phonemes are strongly predictive of lexical decision scores for the most frequent words in the English language (~10,000 words) well beyond frequent rhymes. A good example of this are words like *dog, job,* and *big*, which all have rhyme counts of 1 for the most 2,500 frequent words but reports Levenshtein distances below 1.5 (i.e., average number of insertions, deletions, or transposition operations needed to change one word into another for its top 20 neighbors). They are also among the words with the lowest lexical decision times in the ELP.

When rhyme count features were used in a regression model to predict word decoding times for all words, along with neighborhood effect features, three features related to word frequency, phonographic neighborhood density, and rhymes for the most frequent 2,500 words explained 52% of the overall variance. The word frequency measure explained the greatest amount of the variance (56%) while the phonographic neighborhood density measure explained 41% of the variance and the rhyme count variable explained 3% of the variance. The results showed that words that had faster decoding were words that were more frequent, included more phonographic neighbors, and had more rhymes. Good examples of differences between rhymes and phonographic neighborhood density scores are words such as *lion, happy,* and *noise*, which all have zero counts for rhymes for the most 2,500 frequent words but report phonographic neighborhood densities around 2. These words are also among the words with the lowest lexical decision times. Both lexical decision time analyses indicated that there are strong links between word decoding and word frequency and neighborhood density measures while rhyme measures reported weaker links.

Overall, our analyses find that word frequency measures are the stronger predictor of text readability and word decoding, which has been seen in numerous previous studies. However, the rhyme counts derived from PeRDict explained unique variance beyond that reported by word frequency, as did measures of neighborhood density for the word decoding analyses. The frequency measure was not included in the word decoding analysis that focused on the most frequent words because it was highly multicollinear with a measure of Levenshtein Distance for phonological neighbors.

Our main interest is in comparing the quantitative measures that tap into assessing phonological awareness from PeRDict and ELP. In terms of readability, The PeRDict variable related to the number of rhyme lemmas for the most frequent 1,000 words reported stronger correlations with readability than all ELP variables. Additionally, no ELP neighborhood density variables were included in the final regression model because they explained no unique variance beyond word frequency or rhyme counts. Thus, it seems that rhyme counts are stronger predictors of readability than ELP neighborhood density variables for the readability criterion assessed in this study. It is possible that a different corpus of readability that is not focused on teachers' judgments of text difficulty for K-12 students may report different findings.

The rhyme count features from PeRDict were weaker at predicting word decoding times when compared to their performance in predicting text comprehension. In both analyses, a measure of neighborhood density far outperformed the rhyme measures. A likely explanation is that rhyme counts do not strongly overlap with features that influence adult decoding skills, which is reflected in the ELP lexical decision time responses. In the analysis of the lexical decision times for the most frequent "10,000" words, the Levenshtein distance for phonological neighbors explained 85% of the variance. In the full dataset of words, phonographic neighborhood density scores explained 41% of the lexical decision scores. Both Levenshtein distance and phonographic neighborhood density variables include rhymes in their counts, but rhymes are not the only feature measured. For instance, the neighbors for *cat* include the rhymes *cat* and *rat* as well as non-rhymes including *car* and *cot*. Additionally, the rhymes *combat* and *matte* would not be reflected as a neighbor of *cat.* The aggregation of rhymes along with other types of neighbors in the ELP neighborhood features provides greater coverage of phonological awareness. This is both a strength and a limitation of neighborhood counts.

Specifically, while neighborhood counts are more predictive, researchers interested in quantitatively assessing phonological awareness cannot use these counts to assess the strength of rhymes over non-rhyme neighbors and cannot disaggregate specific types of non-rhyme neighbors from others (i.e., whether the neighbors are based on changes to vowel or consonants in the original word or changes in the initial, medial, or final phonemes in the original word). The aggregated properties found in neighborhood density score likely affect their strength in predicting adult lexical decision times. Specifically, because adults have fuller representations for word's phonological properties, measure that includes rhymes and neighbors are likely stronger predictors. It is possible that children have different decoding strategies than adults and these strategies may depend more on rhymes (although this needs to be tested).

Implications. The results reported in this study have practical implications in terms of automatic assessment of phonological awareness as well as for our understanding of language processing and development. In terms of automatic assessment, the PeRDICT database will allow researchers in academia and industry to assess phonological awareness using more robust natural language processing tools. The inclusion of variables related to perfect rhymes in readability formulas or in purpose-built tools to assess decoding could better assess the difficulty of texts and better match readers with texts that match their ability levels. This could allow for enhanced reading skills development and increased text comprehension.

The results also help us to better understand language processing and development. For instance, the findings from this study indicate that K-12 readers are more likely aided by perfect rhymes when processing text than by neighborhood effects. This effect seems to build on top of word frequency such that a unique amount of reading decoding is based on perfect rhymes. This trend changes slightly for adult decoding metrics where neighborhood effects outperform perfect rhymes, but perfect rhymes do still explain some variance in decoding values.

Conclusion

The analyses reported in this paper assess the potential to use perfect rhymes (as reported by the PeRDict database) to assess text readability and decoding. A main benefit of this approach is that the PeRDict database provides a more fine-grained measurement of word rhymes than the neighborhood effect features reported in the ELP database because those neighborhood effects include both rhymes and non-rhymes.

The results of the study indicate a number of logical next steps for developing additional quantitative measures that tap into phonological awareness and assessing those measures. For instance, neighborhood effect counts could be further disaggregated into non-rhymes with counts for words created by making changes to initial, medial, and final sounds separated by both vowels and consonants. As an example, a word like *love* has rhyme neighbors (e.g., *shove*), non-rhyme neighbors based on initial consonant changes (e.g., *cove*), medial vowel changes (e.g., *live*), and final consonant changes (e.g., *lobe*), and rhymes that are not neighbors (e.g., *above*). These distinctions could be disaggregated and tested in future studies. It is likely that neighborhood effects for rhymes and non-rhymes might explain unique variance in decoding data, providing additional information on language processing and development.

In terms of rhymes, there is a possible confound between rhyme frequencies, rhyme neighborhood density, and word frequencies for rhyme neighbors, which future studies could also disaggregate. Beyond rhymes and neighborhood effects, there are also a variety of other sound and letter features related to syllable structure and phonemes that should be developed to better quantify phonological awareness. These features might include character and syllable entropy (i.e., the certainty of the character and syllable order in words), syllable density (e.g., length of syllables and number of vowel and consonants in syllables), and Levenshtein distances for rhymes. Developing datasets for these features along with code to automatically derive these features from texts would help to move automatic analyses of decoding forward.

The development of more granular non-rhyme counts that tap into phonological awareness along with the rhyme features available in the PeRDict database would provide researchers with greater control over analyses exploring phonological awareness and provide researchers with increased opportunities to understand how the phonological properties of words interact with language learning, development, and processing. Of particular interest would be the assessment of child language production (either spoken or written) to examine links between quantitative measures that tap into phonological awareness and language and writing development as well as studies that use children's language production to assess their phonological awareness (as compared to using standardized tests). Overall, there are several important research avenues still open to investigation in terms of quantitatively measuring features related to phonological awareness.

Crossley, S. A., & Choi, J. S. (in press). Measuring Phonological Complexity Using the Perfect Rhymes Dictionary (PeRDict). *Reading Psychology*.

Open Practices Statement

The PeRDict database is available at https://anonymous.4open.science/r/PeRDict-database-1D68/README.md. All data and materials for all experiments reported in this paper are available at https://anonymous.4open.science/r/PeRDict-Analyses-8E56/README.md.

Crossley, S. A., & Choi, J. S. (in press). Measuring Phonological Complexity Using the Perfect Rhymes Dictionary (PeRDict). *Reading Psychology*.

References

Andrews, S. (1989). Frequency and neighborhood effects on lexical access: Activation or search?. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*(5), 802.

Andrews, S. (1992). Frequency and neighborhood effects on lexical access: Lexical similarity or orthographic redundancy?. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(2), 234.

Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic bulletin & review, 4*(4), 439-461.

Anthony, J. L., Lonigan, C. J., Burgess, S. R., Driscoll, K., Phillips, B. M., & Cantor, B. G. (2002). Structure of preschool phonological sensitivity: Overlapping sensitivity to rhyme, words, syllables, and phonemes. *Journal of experimental child psychology, 82*(1), 65-92.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1996). *CELEX*. Philadelphia: Linguistic Data Consortium.

Ball, E. W., & Blachman, B. A. (1991). Does phoneme awareness training in kindergarten make a difference in early word recognition and developmental spelling?. *Reading research quarterly*, 49-66.

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception & Performance*, *10*, 340-357.

Balota, D.A., Yap, M.J., Cortese, M.J., Hutchison, K.A., Kessler, B., Loftis, B., Neely, J.H., Nelson, D.L., Simpson, G.B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39*, 445-459.

Crossley, S. A., & Choi, J. S. (in press). Measuring Phonological Complexity Using the Perfect Rhymes Dictionary (PeRDict). *Reading Psychology*.

Blachman, B. A. (1984). Relationship of rapid naming ability and language analysis skills to kindergarten and first-grade reading achievement. *Journal of Educational Psychology, 76*(4), 610.

Bradley, L., & Bryant, P. E. (1978). Difficulties in auditory organisation as a possible cause of reading backwardness. *Nature, 271*(5647), 746-747.

Bradley, L., & Bryant, P. E. (1983). Categorizing sounds and learning to read—a causal connection. *Nature, 301*(5899), 419-421.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika, 39*(3/4), 324-345.

Bryant, P. E., MacLean, M., Bradley, L. L., & Crossland, J. (1990). Rhyme and alliteration, phoneme detection, and learning to read. *Developmental psychology, 26*(3), 429.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior research methods, 41*(4), 977-990.

Calfee, R. C., Lindamood, P., & Lindamood, C. (1973). Acoustic-phonetic skills and reading: Kindergarten through twelfth grade. *Journal of Educational Psychology, 64*(3), 293.

Carroll, J. M., Snowling, M. J., Stevenson, J., & Hulme, C. (2003). The development of phonological awareness in preschool children. *Developmental psychology, 39*(5), 913.

Cassady, J. C., Smith, L. L., & Putman, S. M. (2008). Phonological awareness development as a discrete process: Evidence for an integrative model. *Reading Psychology, 29*(6), 508–533.

Catts, H. W., Fey, M. E., Zhang, X., & Tomblin, J. B. (2001). Estimating the risk of future reading difficulties in kindergarten children.

Crossley, S. A., & Choi, J. S. (in press). Measuring Phonological Complexity Using the Perfect Rhymes Dictionary (PeRDict). *Reading Psychology*.

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In Dornic, S. (Ed.), *Attention and performance VI.* New York: Academic Press. (pp. 535-555).

Crossley, S. A., & Kyle, K. (2022). Managing SLA data with NLP tools. In A Berez-Kroeker, B. McDonnell, E. Koller, & L. Collister (Eds). *The Open Handbook of Linguistic Data Management* (pp 411-421). Cambridge, MA: The MIT Press. https://doi.org/10.7551/mitpress/12200.003.0039

Crossley, S. A., Heintz, A., Choi, J., Batchelor, J., & Karimi, M. (2021). The CommonLit Ease of Readability (CLEAR) Corpus. Proceedings of the 14th International Conference on Educational Data Mining (EDM). Paris, France.

Crossley, S. A., Heintz, A., Choi, J., Batchelor, J., Karimi, M., & Malatinszky, A. (in press). A large-scaled corpus for assessing text readability. *Behavior Research Methods*.

Davies, M. (2009-) The Corpus of Contemporary American English (COCA). Available online at https://www.english-corpora.org/coca/.

De Cara, B., & Goswami, U. (2002). Similarity relations among spoken words: The special status of rimes in English. *Behavior Research Methods, Instruments, & Computers, 34*(3), 416-423.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J., Gruber, B., Lafourcade, B., Leitão, P.J., Münkemüller, T., McClean, C., Osborne, P.E., Reineking, B., Schröder, B., Skidmore, A.K., Zurell, D., & Lautenbach, S. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography, 36*(1), 27-46.

Crossley, S. A., & Choi, J. S. (in press). Measuring Phonological Complexity Using the Perfect Rhymes Dictionary (PeRDict). *Reading Psychology*.

Duncan, L. G., Seymour, P. H., & Hill, S. (1997). How important are rhyme and analogy in beginning reading?. *Cognition, 63*(2), 171-208.

Elbro, C., & Jensen, M. N. (2005). Quality of phonological representations, verbal learning, and phoneme awareness in dyslexic and normal readers. *Scandinavian Journal of Psychology, 46*(4), 375-384.

Fox, B., & Routh, D. K. (1975). Analysing spoken language into words, syllables, and phonemes: A developmental study. *Journal of Psycholinguistic Research, 4*(4), 331–342.

Gilhooly, K.J., Logie, R.H. Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation 12*, 395–427 (1980).

Gillon, G. T. (2004). *Phonological awareness: From research to practice*. Guilford Press.

Goswami, U., & Bryant, P. (1990). *Phonological skills and learning to read*. Lawrence Erlbaum Associates, Inc.

Grömping, U. (2006). Relative Importance for Linear Regression in R: The Package relaimpo. *Journal of Statistical Software, 17*(1), 1–27.

Gunning, T.G. (2001) *Creating literacy instruction for all children*. Boston: Allyn & Bacon.

Hipfner-Boucher, K., Milburn, T., Weitzman, E., Greenberg, J., Pelletier, J., & Girolametto, L. (2014). Relationships between preschoolers' oral language and phonological awareness. *First Language, 34*(2), 178-197.

Høien, T., Lundberg, I., Stanovich, K. E., & Bjaalid, I. K. (1995). Components of phonological awareness. *Reading and writing, 7*(2), 171-188.

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength Natural Language Processing in Python*.

Crossley, S. A., & Choi, J. S. (in press). Measuring Phonological Complexity Using the Perfect Rhymes Dictionary (PeRDict). *Reading Psychology*.

Hulme, C. (2002). Phonemes, rimes, and the mechanisms of early reading development. *Journal of Experimental Child Psychology, 82*(1), 58-64.

Katz, L., Brancazio, L., Irwin, J., Katz, S., Magnuson, J., & Whalen, D. H. (2012). What lexical decision and naming tell us about reading. *Reading and writing, 25*(6), 1259-1282.

Kim, M., Crossley, S. A., & Skalicky, S. (2018). Effects of lexical features, textual properties, and individual differences on word processing times during second language reading comprehension. *Reading and Writing, 31*(5), 1155-1180.

Konza, D. (2011). Phonological awareness. *Research into practice, 1*(1), 1-6.

Kucera, H., & Francis, N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of statistical software, 28*, 1-26.

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0. *Behavior research methods, 50*(3), 1030-1046.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady, 10*(8), 707-710.

Lewkowicz, N. K. (1980). Phonemic awareness training: What to teach and how to teach it. *Journal of Educational Psychology, 72*(5), 686.

Lindeman, R. H., Merenda, P. F., & Gold, R. Z. (1980). *Introduction to bivariate and multivariate analysis.*

Lonigan, C. J. (2007). Vocabulary development and the development of phonological awareness

Crossley, S. A., & Choi, J. S. (in press). Measuring Phonological Complexity Using the Perfect Rhymes Dictionary (PeRDict). *Reading Psychology*.

skills in preschool children. In Wagner, R. K., Muse, A. E., & Tannenbaum, K. R. (2007). *Vocabulary acquisition: Implications for reading comprehension*. Guilford Press. (pp. 15-31).

Lonigan, C. J., Anthony, J. L., Phillips, B. M., Purpura, D. J., Wilson, S. B., & McQueen, J. D. (2009). The nature of preschool phonological processing abilities and their relations to vocabulary, general cognitive abilities, and print knowledge. *Journal of educational psychology*, 101(2), 345.

Lukatela, G., & Turvey, M. T. (2000). An evaluation of the two-cycles model of phonology assembly. *Journal of Memory and Language, 42*(2), 183-207.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior research methods, instruments, & computers, 28*(2), 203-208.

Lundberg, I., Frost, J., & Peterson, O. P. (1988). Effects of preschool children. *Reading Research Quarterly, 23*(2).

McDowell, K. D., Lonigan, C. J., & Goldstein, H. (2007). Relations among socioeconomic status, age, and predictors of phonological awareness. *Journal of Speech Language and Hearing Research*, *50*, 1079-1092.

Melby-Lervåg, M., Lyster, S. A. H., & Hulme, C. (2012). Phonological skills and their role in learning to read: a meta-analytic review. *Psychological bulletin, 138*(2), 322.

Metsala, J. L. (1999). Young children's phonological awareness and nonword repetition as a function of vocabulary development. *Journal of educational psychology, 91*(1), 3.

Metsala, J. L., & Walley, A. C. (1998). Spoken vocabulary growth and the segmental restructuring of lexical representations: Precursors to phonemic awareness and early

Crossley, S. A., & Choi, J. S. (in press). Measuring Phonological Complexity Using the Perfect Rhymes Dictionary (PeRDict). *Reading Psychology*.

reading ability. In *J. L. Metsala & L. C. Ehri (Eds.), Word recognition in beginning literacy* (pp. 89–120). Lawrence Erlbaum Associates Publishers.

Morais, J. (2003). Levels of phonological representation in skilled reading and in learning to read. *Reading and Writing, 16*(1), 123-151.

Muter, V., Hulme, C., Snowling, M., & Taylor, S. (1998). Segmentation, not rhyming, predicts early progress in learning to read. *Journal of experimental child psychology, 71*(1), 3-27.

Nithart, C., Demont, E., Metz-Lutz, M. N., Majerus, S., Poncelet, M., & Leybaert, J. (2011). Early contribution of phonological awareness and later influence of phonological memory throughout reading acquisition. *Journal of Research in Reading, 34*(3), 346-363.

Peereman, R., & Content, A. (1995). Neighborhood size effect in naming: Lexical activation or sublexical correspondences?. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(2), 409.

Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific studies of Reading, 18*(1), 22-37.

Pfost, M. (2015). Children's phonological awareness as a predictor of reading and spelling: A systematic review of longitudinal research in German-speaking countries. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 47*(3), 123–138.

Pufpaff, L. A. (2009). A developmental continuum of phonological sensitivity skills. *Psychology in the Schools, 46*(7), 679-691.

Rastle, K., & Coltheart, M. (2006). Is there serial processing in the reading system; and are there local representations? In S. Andrews (Ed.), *From inmarks to ideas: Current issues in lexical processing* (pp. 3-24). Hove, UK: Psychology Press.

Crossley, S. A., & Choi, J. S. (in press). Measuring Phonological Complexity Using the Perfect Rhymes Dictionary (PeRDict). *Reading Psychology*.

Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition, 14*, 191-201.

R Core Team (2020). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Schaefer, B., Fricke, S., Szczerbinski, M., Fox-Boyer, A. V., Stackhouse, J., & Wells, B. (2009). Development of a test battery for assessing phonological awareness in German-speaking children. *Clinical linguistics & phonetics, 23*(6), 404-430.

Schatschneider, C., Fletcher, J. M., Francis, D. J., Carlson, C. D., & Foorman, B. R. (2004). Kindergarten prediction of reading skills: A longitudinal comparative analysis. *Journal of educational psychology, 96*(2), 265.

Share, D. L., Jorm, A. F., Maclean, R., & Matthews, R. (1984). Sources of individual differences in reading acquisition. *Journal of educational Psychology, 76*(6), 1309.

Stanovich, K. E., Cunningham, A. E., & Cramer, B. B. (1984). Assessing phonological awareness in kindergarten children: Issues of task comparability. *Journal of experimental child psychology, 38*(2), 175-190.

Stuart, M. (2005). Phonemic analysis and reading development: Some current issues. Journal of *Research in Reading, 28*(1), 39-49.

Swanson, H. L., & Alexander, J. E. (1997). Cognitive processes as predictors of word recognition and reading comprehension in learning-disabled and skilled readers: Revisiting the specificity hypothesis. *Journal of Educational Psychology, 89*(1), 128–158.

Crossley, S. A., & Choi, J. S. (in press). Measuring Phonological Complexity Using the Perfect Rhymes Dictionary (PeRDict). *Reading Psychology*.

Verhoeven, L., van Leeuwe, J., & Vermeer, A. (2011). Vocabulary growth and reading development across the elementary school years. *Scientific Studies of Reading, 15*(1), 8-25.

Wagner, R. (1988). Causal relations between the development of phonological processing abilities and the acquisition of reading skills: A meta-analysis. *Merrill-Palmer Quarterly 34*, 261-279.

Walley, A. C. (1993). The role of vocabulary development in children′ s spoken word recognition and segmentation ability. *Developmental review, 13*(3), 286-350.

Weber, R. M. (2006). Function words in the prosody of fluent reading. *Journal of Research in Reading, 29*(3), 258-269.

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic bulletin & review, 15*(5), 971-979.

Yopp, H. K. (1988). The validity and reliability of phonemic awareness tests. *Reading research quarterly*, 159-177.

Zipf, G. K. (1935). *The psycho-biology of language*. New York: Houghton-Mifflin.

Table 1
Example data from CMUdict used in PeRDict

Word	Phonemes reported for word					
Nap	N	AE1	P			
Map	M	AE1	P			
Maps	M	AE1	P	S		
Collapse	K	AH0	L	AE1	P	S

Table 2
Number of words and words with rhymes in each adjusted subset

Subset	Number of words	Number of words with rhymes
1,000-word band	994	538
2,500-word band	2,471	1,421
5,000-word band	4,895	2,848
10,000-word band	9,538	5,733
Entire dataset	47,865	30,814

Table 3
Correlations between selected features and reading ease score

Feature	2	3	4	5	6	7
1. Reading ease score	0.509	0.448	0.394	0.293	0.218	0.178
2. SUBTLEXus frequency	1	0.600	0.687	0.416	0.32	0.556
3. Rhymes lemma 1,000		1	0.538	0.699	0.623	0.343
4. Phonographic Neighborhood Density			1	0.404	0.322	0.465
5. Rhymes token 2,500				1	0.603	0.290
6. Rhymes lemma 10,000					1	0.238
7. Phonological neighbor frequency						1

Table 4

Linear model to predict reading ease score

Variable	Relative Importance	Estimate	SE	*t*	*p*
(Intercept)		-0.9564	0.01511	-63.3	< .001
SUBTLEXus frequency	0.530	0.39163	0.01907	20.54	< .001
Rhymes lemma 1,000	0.470	0.23069	0.01902	12.13	< .001

Table 5

Correlations between selected features and lexical decision times for 10,000 most frequent words (tokens)

Feature	2	3	4	5	6	7
1. Lexical decision Time	0.558	-0.133	-0.221	-0.254	-0.295	-0.387
2. Levenshtein Distance phonological neighbors	1	-0.143	-0.182	-0.217	-0.533	-0.679
3. Rhymes token 5,000		1	0.379	0.645	0.18	0.241
4. Rhymes token 1,000			1	0.649	0.179	0.272
5. Rhymes token 2,500				1	0.214	0.319
6. Phonographic Neighborhood Density					1	0.554
7. Phonological neighbors frequency (HAL)						1

Table 6
Linear model to predict lexical decision times for "10,000" most frequent words

Variable	Relative Importance	Estimate	SE	*t*	*p*
(Intercept)		-0.362	0.003	-120.467	0.984
Levenshtein Distance phonological neighbors	0.848	0.146	0.003	48.200	< .001
Rhymes token 1,000	0.061	-0.016	0.004	-4.190	< .001
Rhymes token 2,500	0.091	-0.029	0.004	-7.232	< .001

Table 7

Correlations between selected features and lexical decision times for all words (tokens)

Feature	2	3	4	5	6	7	8	9
1. Lexical decision Time	0.572	-0.107	-0.149	-0.169	-0.195	-0.303	-0.393	-0.644
2. Levenshtein Distance orthographic neighbors	1	-0.091	-0.123	-0.134	-0.156	-0.536	-0.642	-0.389
3. Rhymes token 10,000		1	0.245	0.68	0.419	0.094	0.132	0.176
4. Rhymes token 1,000			1	0.422	0.669	0.157	0.191	0.251
5. Rhymes token 2,500				1	0.68	0.161	0.196	0.253
6. Rhymes token 5,000					1	0.182	0.229	0.294
7. Phonographic Neighborhood Density						1	0.489	0.289
8. Phonological neighbors frequency (HAL)							1	0.442
9. SUBTLEXus frequency								1

Table 8

Linear model to predict all lexical decision times

Variable	Relative Importance	Estimate	SE	*t*	*p*
(Intercept)		-0.093	0.002	-50.523	< .001
SUBTLEXus frequency	0.563	-0.203	0.002	-97.896	< .001
Phonographic Neighborhood Density	0.410	0.154	0.002	76.794	< .001
Rhymes token 2,500	0.027	0.004	0.002	2.332	< .010

---

[i] The PeRDict database is freely available at https://anonymous.4open.science/r/PeRDict-database-595B/README.md

[ii]Studies have indicated that awareness of smaller units of sound at the phoneme level may be better predictors of differences in reading proficiency in English than larger units of sound like rhymes (Hoien, Lundberg, Stanovich, & Bjaalid, 1995; Muter, et al., 1998; Nation & Hulme, 1997).

[iii] See Yarkoni et al. (2008) for counterevidence to this general finding.

[iv] The website that hosted the dataset reported in de Cara and Goswami (2002) is no longer active.

[v] The Python script used for this task is available at https://anonymous.4open.science/r/PeRDict-Analyses-72DA/README.md

[vi] All R scripts and data for the analyses in this paper are available at https://anonymous.4open.science/r/PeRDict-Analyses-72DA/README.md