Behav Res (2018) 50:1030–1046
DOI 10.3758/s13428-017-0924-4

The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0

Kristopher Kyle[1] · **Scott Crossley**[2] · **Cynthia Berger**[2]

Published online: 11 July 2017

Abstract This study introduces the second release of the Tool for the Automatic Analysis of Lexical Sophistication (TAALES 2.0), a freely available and easy-to-use text analysis tool. TAALES 2.0 is housed on a user's hard drive (allowing for secure data processing) and is available on most operating systems (Windows, Mac, and Linux). TAALES 2.0 adds 316 indices to the original tool. These indices are related to word frequency, word range, *n*-gram frequency, *n*-gram range, *n*-gram strength of association, contextual distinctiveness, word recognition norms, semantic network, and word neighbors. In this study, we validated TAALES 2.0 by investigating whether its indices could be used to model both holistic scores of lexical proficiency in free writes and word choice scores in narrative essays. The results indicated that the TAALES 2.0 indices could be used to explain 58% of the variance in lexical proficiency scores and 32% of the variance in word-choice scores. Newly added TAALES 2.0 indices, including those related to *n*-gram association strength, word neighborhood, and word recognition norms, featured heavily in these predictor models, suggesting that TAALES 2.0 represents a substantial upgrade.

Keywords Lexical sophistication · Natural language processing · Writing quality

Lexical sophistication is an important consideration in fields such as educational psychology, cognitive science, and artificial intelligence, where text complexity, learning trends, and language production are important areas of study. However, lexical sophistication can be measured in a number of different ways, with perhaps the most common measure being word frequency. Word frequency measures calculate how frequently a word occurs in general usage, as measured by a representative corpus such as the British National Corpus (BNC; BNC Consortium, 2007), the Corpus of Contemporary American English (COCA; Davies, 2010), or SUBTLEXus (Brysbaert & New, 2009). Word frequency has been shown to be a strong predictor of lexical-response and word-naming times (Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; Forster & Chambers, 1973; Frederiksen & Kroll, 1976), as well as to be strongly correlated with a number of related developmental constructs, such as writing and speaking proficiency (Kyle & Crossley, 2015; Laufer & Nation, 1995; McNamara, Crossley, Roscoe, Allen, & Dai, 2015), and text complexity considerations, such as reading difficulty (Crossley, Dufty, McCarthy, & McNamara, 2007; Nation, 2006). Recent research, however, has suggested that features other than lexical frequency may explain lexical knowledge and development better than word frequency does (Adelman, Brown, & Quesada, 2006; Crossley, Salsbury, & McNamara, 2012; Johns & Jones, 2008; Kyle & Crossley, 2015; McDonald & Shillcock, 2001).

In this article, we introduce and test the reliability of the second version of the Tool for the Analysis of Lexical Sophistication (TAALES 2.0). TAALES 1.0 (Kyle & Crossley, 2015) was developed to provide researchers with a freely available tool that would automatically calculate a variety of classic and new indices of lexical sophistication. It included lexical features related to word frequency, word range, *n*-gram frequency, academic language, and psycholinguistic word properties. The second iteration of TAALES increases the breadth and depth of the available indices reported by TAALES by expanding the word frequency, word range, and *n*-gram frequency indices, and by adding indices related to word recognition norms,

✉ Kristopher Kyle
kristopherkyle1@gmail.com

[1] University of Hawaii at Manoa, Honolulu, HI, USA

[2] Georgia State University, Atlanta, Georgia

contextual distinctiveness, word neighborhood, semantic network, *n*-gram range, and *n*-gram strength of association. To help validate the indices reported in TAALES 2.0, we present two studies. In the first study, indices of lexical sophistication are used to model human judgments of lexical proficiency in free writes. In the second study, the TAALES 2.0 indices are used to model word-choice ratings in narrative essays.

TAALES 1.0

Given the importance of lexical sophistication in a number of fields as well as the relative difficulty of accessing methods for the assessment of lexical sophistication beyond word frequency and/or type–token ratios, we developed TAALES 1.0 (Kyle & Crossley, 2015). The most recent version of TAALES 1.0 (version 1.4) included 104 indices, related to word frequency, word range, *n*-gram frequency, academic language, and psycholinguistic word information. TAALES versions 1.0–1.4 have been used in a variety of domains, including the assessment of written lexical proficiency, speaking proficiency, and writing quality (Allen, Crossley, & McNamara, 2015; Allen & McNamara, 2015; Jung, Crossley, & McNamara, 2015; Kyle & Crossley, 2015); modeling lexical development (Crossley, Kyle, & Salsbury, 2016); identifying satire in product reviews (Skalicky & Crossley, 2015); and indexing humor in academic writing (Skalicky, Berger, Crossley, & McNamara, 2016). Below we provide a brief description of the constructs covered in TAALES 1.4 (see Kyle & Crossley, 2015, for a comprehensive treatment).

Word frequency

Word frequency refers to the number of times a word occurs in a corpus of texts. Words that are less frequent in a reference corpus (e.g., *edifice*, *cuisine*, *egregious*) are considered more sophisticated than words that occur frequently (e.g., *building*, *food*, *bad*). A great deal of research has demonstrated the relationship between the frequency of lexical items in normal language use and lexical sophistication. Reading research has demonstrated that texts that include less frequent lexical items tend to be considered more difficult (Crossley et al., 2007; Nation, 2006). Writing research has indicated that essays that include less frequent lexical items tend to be considered of higher quality (Guo, Crossley, & McNamara, 2013; Laufer & Nation, 1995; McNamara et al., 2015), and similar findings have been observed with regard to written lexical proficiency and speaking proficiency (Kyle & Crossley, 2015). TAALES 1.4 includes 36 frequency indices derived from the BNC (BNC Consortium, 2007), the Brown verbal frequency list (Brown, 1984; Svartvik & Quirk, 1980), the Kučera–Francis written frequency list (Kučera & Francis, 1967), SUBTLEXus (Brysbaert & New, 2009), and the Thorndike–Lorge written frequency list (Thorndike & Lorge, 1944).

Word range

Range refers to the number of texts in a corpus in which a particular item occurs. Although a robust relationship between frequency and language proficiency has been established, word frequency values may be inflated due to a high occurrence of a technical word in a small set of documents in a given corpus that may otherwise be extremely infrequent in general language usage. Range norms help control for this inflation, and may provide a better approximation of an individual's exposure to a particular word. In the written portion of the BNC, for example, the words *next*, *four*, and *cent* have similar frequencies (approximately 381 times per million words). The words *next* and *four* also occur in a wide range of texts (approximately 90% of the texts in the written BNC), but the word *cent* occurs in much fewer texts (approximately 50%). This suggests that despite similar frequencies, *next* and *four* may be more likely to be encountered than *cent*. Range indices have recently been used to model writing quality (Kyle & Crossley, 2016), speaking proficiency (Kyle & Crossley, 2015), lexical proficiency (Kyle & Crossley, 2015), and to explain variance in lexical-decision times (Adelman et al., 2006). Generally, words that occur in fewer contexts are considered more sophisticated. Accordingly, language samples that on average include words with a more restricted range tend to earn higher quality/proficiency scores. TAALES 1.4 includes 18 range indices derived from the BNC, Kučera–Francis written frequencies, and SUBTLEXus.

Academic language

Learning academic language is an important part of academic socialization (Hyland, 2009). Academic language includes words and phrases that occur frequently in academic contexts, but infrequently in general language use. Perhaps the most influential list of academic language is Coxhead's (2000) academic word list (AWL), which has been integrated into English for academic purposes (EAP) research and pedagogy (Coxhead, 2011). A similar list of academic multiword formulas (the Academic Formulas List [AFL]) was developed by Simpson-Vlach and Ellis (2010). A higher proportion of academic language in a text would lead to a more sophisticated text. However, the few studies that have employed AFL and AWL indices have failed to find a relationship between the use of academic language and writing proficiency/lexical sophistication (Kyle & Crossley, 2015). TAALES 1.4 included 15 indices related to academic language derived from the AWL and the AFL.

N-gram frequency

The individual word has a long history as the unit of investigation in vocabulary and lexical development studies (e.g.,

Laufer & Nation, 1995; Nation, 2006). Recent research, however, has begun to highlight the importance of multiword units (Biber, Conrad, & Cortes, 2004; O'Donnell, Römer, & Ellis, 2013) in language development. *N*-gram frequencies have recently been used to model writing quality (Bestgen & Granger, 2014; Crossley, Cai, & McNamara, 2012; Kyle & Crossley, 2016), speaking proficiency (Kyle & Crossley, 2015), and to model scores of lexical proficiency (Kyle & Crossley, 2015). *N*-grams such as *the end of*, *out of the*, and *a lot of* occur frequently, whereas *n*-grams such as *now not only*, *time some of*, and *is about being* occur much less frequently. Across these assessment contexts, *n*-gram frequencies have generally been positively correlated with proficiency/quality scores (cf. Crossley, Cai, et al. 2012), suggesting that an indicator of linguistic development is the knowledge of how words tend to be combined. TAALES 1.4 includes 12 indices related to *n*-gram frequency derived from the BNC.

Psycholinguistic word information

The psycholinguistic properties of words have been of interest to psycholinguists, cognitive scientists, and second language acquisition researchers for some time (Coltheart, 1981; Crossley et al., 2016; Crossley, Weston, Sullivan, & McNamara, 2011; Toglia & Battig, 1978). Word properties such as concreteness, familiarity, meaningfulness, imageability, and age of acquisition have been used to model writing quality scores (Crossley & McNamara, 2011; Guo et al., 2013), speaking proficiency (Crossley & McNamara, 2013; Kyle & Crossley, 2015), lexical proficiency (Crossley, Salsbury, & McNamara, 2012; Kyle & Crossley, 2015), lexical-decision times (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012), and word associations (Altarriba, Bauer, & Benvenuto, 1999). TAALES 1.4 includes 21 indices related to concreteness, familiarity, meaningfulness, and age of acquisition derived from the MRC database (Coltheart, 1981), Brysbaert, Warriner, and Kuperman (2014), and Kuperman et al. (2012).

TAALES 2.0

Like its predecessor, TAALES 2.0 is freely available,[1] easy to use, and compatible with most operating systems (Windows, Mac, and Linux). It is written in Python, is accessed via an intuitive graphical user interface (GUI; see Fig. 1), and requires no programming knowledge to operate. Unlike Web-based tools, TAALES 2.0 is housed on the user's hard drive, which allows users to process data securely and without the need for an Internet connection. TAALES 2.0 includes all of the indices found in TAALES 1.4 and adds over 300 additional indices, related to word and *n*-gram frequency and range, *n*-gram strength of association, contextual distinctiveness, word recognition norms, semantic network, and word neighbors. Each of these is described below. For an overview of the indices included in TAALES 1.4 and TAALES 2.0, see Table 1. For instructions regarding the use of TAALES 2.0, see the user manual, which is available as supplementary material at www.kristopherkyle.com/supplementary-materials.html.

Word and *n*-gram frequency and range

Indices related to word and *n*-gram frequency and range have been shown to be important predictors of a number of constructs related to language development. One factor that may affect the accuracy of word frequency and range indices is the reference corpora from which norms are derived. These factors include, but are not limited to, mode, region, and purpose (i.e., register; Biber, Conrad, Reppen, et al., 2004; Hyland, 2009). Thus, TAALES 2.0 includes frequency and range norms for words and *n*-grams that are register-specific and derived from the Corpus of Contemporary American English (COCA; Davies, 2009). COCA comprises texts collected between 1990 and 2015 that represent five registers (academic, fiction, magazine, news, and spoken). In total, COCA includes approximately 520 million words. TAALES 2.0 includes six word frequency and six word range indices for each of the five COCA registers (30 total for each), 24 *n*-gram frequency (bigram and trigram) indices for each COCA register (120 total), and four *n*-gram range indices for each COCA register (20 total). Additionally, TAALES 2.0 adds two word frequency indices from the 131-million-word Hyperspace Analogue to Language (HAL) corpus (Lund & Burgess, 1996), compiled from Internet news groups.[2]

Age of exposure

TAALES 2.0 includes recently developed age of exposure (AoE) indices (Dascalu, McNamara, Crossley, & Trausan-Matu, 2016). These indices are based on computational models that estimate a word's complexity on the basis of co-occurrence data and a word's links to relevant sematic concepts within large corpora. Using latent Dirichlet allocation (LDA), which computationally infers underlying topics

[1] TAALES 2.0 is freely available at www.kristopherkyle.com/taales.html under a Creative Commons Attribution-NonCommercial-ShareAlike International license. All of the included databases are free for noncommercial research purposes at the time of writing, but they may fall under a use license other than that of TAALES. Researchers should check each database source (available in the supplementary material document entitled TAALES_2.0_Index_Guide.xlsx, which is available at www.kristopherkyle.com/supplementary-materials.html) to determine whether their project falls within the guidelines and/or license for each.

[2] TAALES does not provide access to COCA or the HAL corpus. The COCA frequency and collocational data is available at www.wordfrequency.info. Licenses for accessing the texts that comprise COCA are available at www.corpusdata.org.

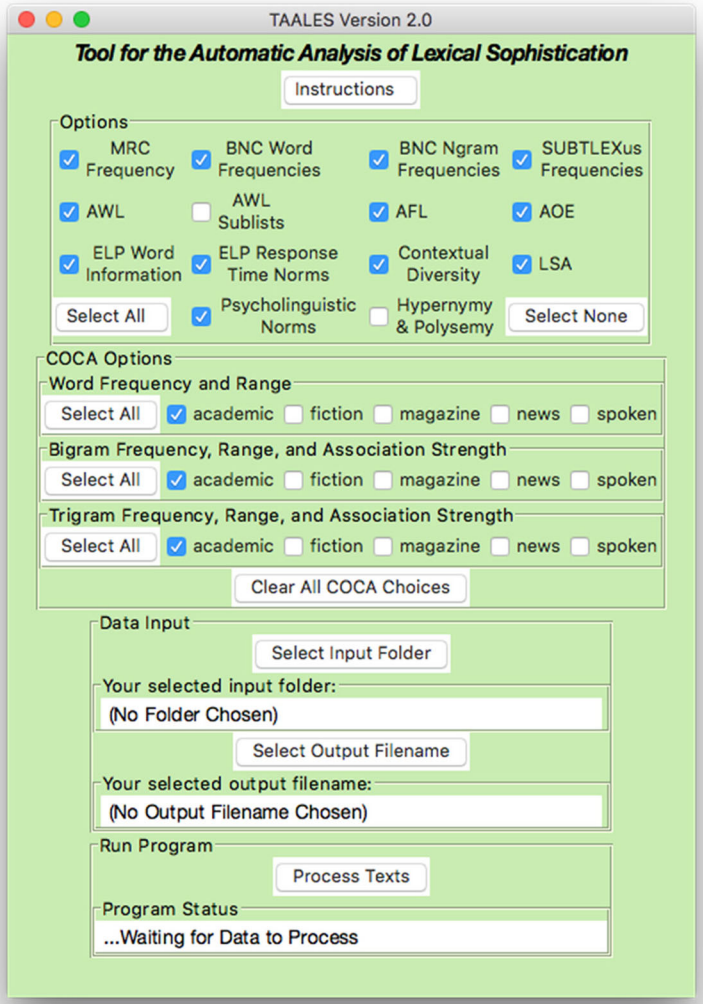

Fig. 1 TAALES 2.0 graphical user interface

through a generative probabilistic process, Dascalu et al. developed measures of AoE values for the words found in the Touchstone Applied Science Associates (TASA) corpus,[3] which contains 13 grade-level textbooks in the United States (Landauer, Foltz, & Laham, 1998). Validations of the AoE values indicate that they are strongly related to human ratings of age of acquisition, word frequency, entropy, and human lexical-response latencies.

Word recognition norms

Word recognition scores report the average response latencies, standard deviations, and accuracies for a given word when used as a stimulus in lexical-decision and word-naming tasks. Lexical-decision latencies (i.e., response times) measure the time it takes participants to decide whether a word is a real word in English or not, whereas word-naming latencies measure the time it takes participants to begin reading a word aloud. These norms may reflect the ease or difficulty of processing a given word (Balota et al., 2004; Forster & Chambers, 1973; Frederiksen & Kroll, 1976). TAALES 2.0 calculates eight indices based on lexical-decision (LD) and word-naming (WN) behavioral norms obtained from The English Lexicon Project (ELP), a large publicly available psycholinguistic dataset (Balota et al., 2007). The ELP includes LD and WN task response latencies, standard deviations, and accuracies collected from 816 native English-speaking subjects. Word recognition norms were calculated in response to 40,481 real words (and an additional 40,481 nonwords for the LD task).

[3] TAALES also does not provide access to the TASA corpus. Rather, it reports AoE indices derived from that corpus for individual words.

Table 1 Comparison of indices included in TAALES 1.4 and TAALES 2.0

	Indices		Corpora/Databases Represented		
Index Type	1.4	2.0	1.4	2.0	Key Changes in 2.0
Word frequency	36	**68**	5	**11**	Register-specific COCA frequency norms added
Word range	18	**48**	3	**8**	Register-specific COCA range norms added
Psycholinguistic word information	19	14	2	2	Simplified concreteness indices
Age of acquisition/exposure	3	**7**	1	**2**	Corpus-based indices added
Academic words	15	15	2	2	n/a
Contextual distinctiveness	n/a	**8**	n/a	**5**	Indices related to contextual distinctiveness added
Word recognition norms	n/a	**8**	n/a	**1**	Lexical-decision and word-naming indices added
Semantic network	n/a	**14**	n/a	**1**	Polysemy and hypernymy indices added
N-gram frequency	12	**132**	1	**6**	Register-specific COCA frequency norms added
N-gram range	n/a	**20**	n/a	**5**	Register-specific COCA range norms added
N-gram strength of association	n/a	**75**	n/a	**5**	Variety of strength-of-association norms added
Word neighbors	n/a	**14**	n/a	**1**	Orthographic and phonologic neighbor and neighborhood indices added
Other	n/a	**5**	n/a	**2**	Character bigram and age of exposure indices added
Total	**103**	**424**			

Contextual distinctiveness

Contextual distinctiveness measures the diversity of contexts in which a word is encountered. The constraints context puts on a word's meaning may contribute to a more psychologically valid explanation of the word frequency effect than frequency of isolated occurrence alone (Adelman et al., 2006; Brysbaert & New, 2009; McDonald & Shillcock, 2001). Such constraints have been found to predict spoken lexical proficiency in L1 and L2 speech samples (Berger, Crossley, & Kyle, 2017). TAALES includes a number of different techniques for measuring contextual distinctiveness, ranging from free association norms to corpus-driven statistical approaches. These techniques are described below.

Free association norms

One approach to operationalizing contextual distinctiveness is to observe the number of other words commonly associated with a word, such that words with a greater number of associations are assumed to be less contextually distinct. Such information is available from free word association tasks, in which participants are given a stimulus word and asked to produce the first word (or word) that comes to mind.

Two sources of existing L1 word association norms are the Edinburgh Associative Thesaurus (EAT; Kiss, Armstrong, Milroy, & Piper, 1973) and the University of South Florida (USF; Nelson, McEvoy, & Schreiber, 2004) norms. Among other data, the EAT norms include the number of responses a given word receives when used as a stimulus in a written free association task. For example, the word *worry* elicits 65 different response types, whereas the word *husband* elicits only 15. The USF norms report the number of stimuli words that result in production of a given word as an associate in a free association task. Words elicited by a greater range of stimuli are considered more likely to come to mind in response to a variety of cues (Nelson et al., 2004). For example, the word *love* is produced in response to 181 different stimulus words, whereas a less contextually distinct word, such as *bride*, is produced in response to just six stimuli. TAALES 2.0 includes three indices related to free association norms taken from EAT and USF.

Corpus-driven approaches

Other approaches to measuring contextual distinctiveness are based on statistical regularities observed in large reference corpora. One such approach takes a lexical perspective and measures the probability of a given word statistically co-occurring with others words in general language usage (McDonald & Shillcock, 2001). For example, a word like *today* is found within a variety of lexical contexts (i.e., *today* co-occurs with many other words) and is thus less contextually distinct than a word like *lone*, which is less likely to co-occur with other words. Meanwhile, a semantic corpus-based approach to contextual distinctiveness (e.g., Hoffman, Lambon Ralph, & Rogers, 2013) observes the variety of semantic contexts in which a word occurs. The assumption underlying a semantic approach to contextual distinctiveness is that a word occurring in a variety of semantic contexts (e.g.,

one) is more semantically ambiguous and thus less contextually distinct than a word occurring in constrained semantic contexts (e.g., *vibe*). TAALES includes two indices related to corpus derived contextual distinctiveness, including semantic distinctiveness, as reported by Hoffman et al., and the McDonald co-occurrence probability (McDonald & Shillcock, 2001).

Word neighborhood

Word neighborhood refers to the words that share orthographic, phonographic, and/or phonological similarities with a particular word, all of which are correlated with one another (Peereman & Content, 1997). The size and characteristics of a word's neighborhood have been shown to contribute to explaining variance in word-naming and recognition tasks (Adelman & Brown, 2007; Andrews, 1989; Balota et al., 2004; Coltheart, Davelaar, Jonasson, & Besner, 1977; Grainger, 1990; M. Yates, 2005; M. Yates, Locker, & Simpson, 2004). TAALES includes 14 indices related to word neighborhood information derived from the ELP (Balota et al., 2007). These are discussed briefly below.

Orthographic neighbors An orthographic neighbor (Coltheart et al., 1977) is a real word that is formed by changing just one letter in the original word. For example, the word *cat* has 18 orthographic neighbors, including *cab*, *cap*, *car*, *oat*, *sat*, and so forth.

Phonographic neighbors Phonographic neighbors differ in one letter and one phoneme. For example, whereas *stove* and *shove* are only orthographic neighbors, *stone* and *stove* are also phonographic neighbors (Adelman & Brown, 2007).

Phonological neighbors Phonological neighbors differ by one phoneme, regardless of their orthography. For example, the word *geese* has seven phonological neighbors: *cease*, *lease*, *niece*, *peace*, *gas*, *goose*, and *guess* (M. Yates, 2009).

Semantic networks

A *semantic network* refers to the way that word forms are semantically related. Two key areas of semantic networks are polysemy and hypernymy. Both polysemy and hypernymy have been shown to be related to lexical development (Crossley, Salsbury, & McNamara, 2009, 2010) and L2 writing proficiency (Guo et al., 2013; Reynolds, 1995).

Polysemy *Polysemy* refers to the number of related senses (i.e., meanings) a particular word form has. Words such as *make* and *give* have more senses than words such as *construct* and *deliver*. Research has suggested that as learners develop, they tend to use words with fewer senses (Crossley et al., 2010). Furthermore, research has demonstrated that polysemy scores are negatively correlated with L2 writing quality (e.g., Guo et al., 2013). TAALES 2.0 calculates polysemy values for all content words and for nouns, verbs, adjectives, and adverbs (five total indices). Polysemy scores represent the number of senses a word form has according to WordNet (Fellbaum, 1998).

Hypernymy *Hypernymy* refers to the number of superordinate terms a particular word has. A word such as *animal* has but a few superordinate terms, whereas words such as *greyhound*, *stag*, and *whitefish* have many. Research has suggested that as individuals develop, they tend to have access to words with more superordinate terms (i.e., words that are more specific; Crossley et al., 2009). Additionally, hypernymy ratings have been shown to be positively correlated with L2 writing quality (Guo et al., 2013). TAALES 2.0 includes nine indices related to hypernymy for nouns, verbs, and a combination of nouns and verbs. One issue relating to the operationalization of hypernymy is that different senses of a particular word often have different superordinate terms (and different numbers of superordinate forms). Thus, TAALES 2.0 includes three versions of each hypernymy index, such that the first version comprises hypernymy values for the most frequent sense and path, the second comprises the average value for all senses (but the most frequent path for each), and the third version comprises the average value for all senses and all paths.

N-gram strength of association

Strength-of-association norms measure the conditional probability that words will occur together. Strength-of-association norms are related to *n*-gram frequency norms but control for the relative frequencies of the words that comprise *n*-grams by measuring the conditional probability of word co-occurrence. Such norms can show that the words in bigrams such as *optimistic about* are more strongly related than the ones in *and the* and *in the*. Bigram association strength has been shown to be positively correlated with L2 writing quality and longitudinal writing development (Bestgen & Granger, 2014). TAALES 2.0 includes 75 strength-of-association norms, covering both bigrams (25 indices) and trigrams (50 indices). These measures are described below and presented in Table 2. Two types of trigram indices are computed, such that the first word is considered Item 1 and the following bigram is considered Item 2, or the first bigram is considered Item 1 and the third word is considered Item 2.

Mutual information Mutual information (MI) scores represent the joint probability that two items will co-occur. Studies in corpus linguistics have suggested that MI scores tend to inflate the importance of low-frequency items (e.g., bigrams

Table 2 2×2 contingency table used in the calculation of strength-of-association norms

	Item 2	Not Item 2	Totals
Item 1	**a**	**b**	**a** + **b** = frequency of Item 1
Not Item 1	**c**	**d**	**c** + **d** = combinations that are not Item 1 and Item 2
Totals	**a** + **c** frequency of Item 2	**b** + **d**	(**a** + **b**) + (**c** + **d**) = *N* (total number of *n*-gram tokens in the corpus)

that consist of lower-frequency words tend to earn higher MI scores; Evert, 2005). *N*-grams such as *spina bifida* and *lingua franca* earn high MI scores, whereas *n*-grams such as *an a* and *great the* earn low MI scores. MI is calculated as the (logarithm) of the observed co-occurrence of two items divided by the expected co-occurrence of two items:

$$MI = \log\left(\frac{\text{observed}}{\text{expected}}\right) = \log\left(\frac{a}{\frac{(a+b)*(a+c)}{N}}\right).$$

Mutual information squared Mutual information squared (MI^2) scores are a variant of MI scores that attempt to mitigate the emphasis of low-frequency items (Evert, 2005). *N*-grams such as *twentieth century* and *stainless steel* earn high MI^2 scores, whereas *n*-grams such as *the all* and *some and* earn low MI^2 scores. MI^2 is calculated as the (logarithm) of the observed co-occurrence of two items (squared) divided by the expected co-occurrence of two items:

$$MI = \log\left[\frac{\text{observed}^2}{\text{expected}}\right] = \log\left(\frac{a^2}{\frac{(a+b)*(a+c)}{N}}\right).$$

T Like MI scores, T scores represent the joint probability that two items will co-occur. Although MI scores tend to emphasize infrequent items, T scores tend to emphasize frequent items (Evert, 2005). *N*-grams such as *of the* and *from the* earn high T scores, whereas *n*-grams such as *the between* and *the who* earn low T scores. T is calculated as the observed frequency minus the expected frequency, divided by the square root of the observed frequency: $T = \frac{\text{observed}-\text{expected}}{\sqrt{\text{observed}}} = \frac{\text{observed}-\text{expected}}{\sqrt{\text{observed}}}$.

Delta P Delta *P* scores represent the probability of an outcome (i.e., a particular word) based on a cue (i.e., another word). Delta P scores are directional, meaning that word order affects the score, unlike MI, MI^2, and T scores. Delta P is calculated via the following formula: delta P = P(O | C) – P(O | –C); that is, delta P is the probability of an outcome given a cue minus the probability of an outcome without the cue. *N*-grams such as *preformatted table* and *pursuant to* earn high delta P scores, whereas *n*-grams such as *would the* and *must the* earn low delta P scores. With reference to Table 2, we calculate delta P with the second item as the outcome and the first item as the cue via:

$$\text{delta } P = \left(\frac{\text{a}}{\text{a}+\text{b}}\right) - \left(\frac{\text{c}}{\text{c}+d}\right).$$

Approximate collexeme strength Collexeme strength scores (Gries, Hampe, & Schönefeld, 2005) represent the joint probability that two items will co-occur. Collexeme strength is calculated using an exact test and does not include normal distribution as an assumption. For these reasons, it has been argued to be superior to other association strength indices such as MI and T (Gries et al., 2005). Collexeme strength is calculated by taking the negative logarithm of the Fisher–Yates exact test (Fisher, 1934; F. Yates, 1934), which is calculated as:

$$p_{\text{observed distribution}} = \frac{\left(\frac{a+c}{a}\right)*\left(\frac{b+d}{b}\right)}{\frac{N}{a+b}} + \Sigma\, p_{\text{all more extreme distributions}}.$$

Although the use of an exact test has some benefits, in practical applications with large corpora (such as COCA), decimal rounding causes particularly attracted or repelled bigram items to equal 1 or 0, respectively. A solution is to approximate collexeme strength by multiplying the delta P value by the frequency of Item 1:

$$\text{approximate collexeme strength} = \left(\left(\frac{a}{a+b}\right) - \left(\frac{c}{c+d}\right)\right)*(a+b).$$

This approximation is reportedly strongly correlated (r = .950) with collexeme strength (Gries, personal communication, December 19, 2014). *N*-grams such as *for example* and

would be earn high approximate collexeme strength scores, whereas *n*-grams such as *not the* and *more the* earn low approximate collexeme strength scores.

Present studies

In these studies, we validate TAALES 2.0 by investigating whether TAALES 2.0 indices can be used to predict holistic scores of lexical proficiency in L1 and L2 writing samples, and analytic word-choice scores in L1 essays. The research questions that guide these validation studies are

1. What is the relationship between the indices of lexical sophistication included in TAALES 2.0 and holistic scores of written lexical proficiency?
2. What is the relationship between the indices of lexical sophistication included in TAALES 2.0 and word-choice scores for narrative essays?

Method

Corpora

Lexical proficiency corpus The lexical proficiency corpus comprises free writes written by L1 and L2 English users reported by Crossley, Salsbury, McNamara, and Jarvis (2011). It includes 180 free writes written by L2 English learners enrolled in an English for academic purposes program at a university in the US. These texts were stratified to include equal numbers of texts from individuals with beginning (n = 60), intermediate (n = 60), and advanced (n = 60) English proficiency, based on institutional TOEFL scores. These samples were augmented with 60 unstructured writing samples from undergraduate native speakers leading to a total corpus of n = 240. The writing samples were evaluated by expert raters who used a holistic rubric related to lexical proficiency. Interrater reliability was acceptable (r = .796). The corpus has been used in a number of studies to explore the nature of lexical proficiency. Crossley, Salsbury, et al. (2011), for example, found that indices related to lexical diversity, word hypernymy, and content word frequency explained 44% of the variance in lexical proficiency scores. Texts that had higher lexical diversity and included words with fewer hypernyms and lower-frequency content words tended to earn higher scores. In a follow-up study, Kyle and Crossley (2015) found that indices related to bigram and trigram frequency, word range, familiarity, and meaningfulness scores explained 51.7% of the variance in lexical proficiency scores. Texts that included bigrams and trigrams that are less frequent and words that are less frequent, familiar, and meaningful tended to earn higher scores.

Word-choice corpus The word-choice corpus comprises 716 narrative essays written by 10th graders in the United States that predominately speak English as an L1. The corpus was collected as part of the Automated Student Assessment Prize (ASAP) and is described in Shermis and Hamner (2013). Essays were scored using a six trait analytic rubric by at least two raters. For this study, we used analytic ratings related to word choice. The analytic rating for word choice indicates that essays that include "accurate, strong, specific words" would be scored higher for word choice than essays that include "general, vague words" and/or "an extremely limited range of words," which would be scored lower. Interrater reliability for the word-choice ratings was moderate (Kappa = .482), whereas 98.2% of ratings were either exact or adjacent matches. To our knowledge, the word-choice scores have not been used in previous studies. Shermis and Hamner reported on a shared task in which participants attempt to automatically predict overall essay score (which was calculated on the basis of all six analytic traits). Fully featured models (i.e., models that include predictors related to fluency, lexical sophistication, cohesion, and syntactic complexity) explained between 40% and 52% of the variance in essay scores.

TAALES 2.0 indices

All TAALES 2.0 indices related to word frequency, word range, psycholinguistic word information, age of exposure, academic language, contextual distinctiveness, word recognition norms, semantic network, *n*-gram frequency, *n*-gram range, *n*-gram strength of association, and word neighbors were used for the analysis. All TAALES indices are normed by text length. Any item in the text that is not represented in a particular index database (e.g., rare words and misspellings) are not counted toward text length. Some TAALES databases, such as Brysbaert et al.'s (2014) concreteness norms are based on word lemmas, whereas others, such as Balota et al.'s (2007) lexical-decision norms, are based on raw (nonlemmatized) words. Furthermore, some databases/corpora count contractions (e.g., *can't* and *won't*) as two tokens (e.g., *ca* and *n't*), whereas others count them as one. TAALES is sensitive to each of these differences and tokenizes and/or lemmatizes the source texts as necessary. An Index Guide is provided as supplementary material at www.kristopherkyle.com/supplementary-materials.html. The document provides in depth information regarding each index, including database sources and lemmatization information (among other pertinent information). It should also be noted that some databases are larger than others, which may affect index coverage. The MRC concreteness index, for example, is based on a database of concreteness ratings for 4,292 lemmas (Paivio, Yuille, & Madigan, 1968; Spreen & Schulz, 1966), and Brysbaert et al.'s (2014) database includes ratings for 40,000 lemmas. In addition to index

scores, TAALES also provides optional output that indicates the number of text items that are covered by each database providing word coverage for each text for each index.

Statistical analyses

To investigate the relationship between indices of lexical sophistication and human judgments of lexical proficiency in L1 and L2 free writes and word choice in L1 narrative essays, multiple regressions models were developed. For each study, all TAALES 2.0 indices were checked for normality using histograms. Any index that was not normally distributed was removed from further consideration.[4] We then set a correlation threshold of $r = .100$, which represents the lower bound of a meaningful correlation (Cohen, 1988), and our alpha level at $p = .001$ (to control for Type I errors). Any index that did not reach both thresholds was removed from further consideration. We then checked for multicollinearity, which can lead to exaggerated models. Any indices that were strongly correlated ($r = .700$) were flagged for further analysis. In each collinear group, only the index with the strongest correlation with the criteria variable was kept.[5] The remaining indices were then entered into a stepwise regression that used the Akaike information criterion (AIC) method (Akaike, 1974). If any of the indices in the model demonstrated suppression (i.e., their beta weights had switched signs), those indices were removed and the regression was rerun. This process was repeated until the model included no suppressed variables. Finally, a follow-up tenfold forced entry linear regression was conducted using the indices included in the final model to ensure that the model was consistent across the dataset.

Results

Study 1: Lexical proficiency

To validate the indices of lexical sophistication included in TAALES 2.0, 421 indices were used to model the variance in holistic scores of lexical proficiency in essays. Twenty-eight of the indices violated the assumption of normality and were removed from further consideration. Furthermore, 285 of the remaining 393 variables did not reach the minimum correlation threshold of $r \geq .100$ and $p < .001$ and also were removed from further consideration. Of the remaining 108 variables, 84 were removed due to multicollinearity. The remaining 24 variables (see Table 3) were entered into a tenfold stepwise regression. The initial model included two variables with switched signs, which were subsequently removed. The final model, which included ten variables, explained 58.0% ($R^2 = .580$) of the variance in holistic lexical proficiency scores. This model was significant, $F(10, 229)$, $p < .001$. When the model was cross-validated, it explained 56.4% of variance ($R^2 = .564$), suggesting that the model is stable across the dataset. The model included indices related to association strength, *n*-gram proportion scores, range scores, lexical-decision and word-naming response times, age of exposure, word hypernymy and polysemy, and word frequency. The results indicated that texts rated as being more lexically proficiency contained more sophisticated lexical features. Table 4 presents a summary of the regression model.

Study 2: L1 word choice

To validate the indices of lexical sophistication included in TAALES 2.0, 421 indices were used to model the variance in analytic word choice scores in L1 essays. Fourteen indices violated the assumption of normality and were removed from further consideration. One hundred twenty-eight of the remaining 407 variables did not reach the minimum correlation thresholds of $r \geq .100$ and $p < .001$, and were also removed from further consideration. Of the remaining 279 variables, 233 were removed due to multicollinearity. The remaining 46 variables (see Table 5) were entered into a tenfold stepwise regression. The initial model included 11 variables with switched signs, which were subsequently removed. Four additional indices were removed due to suppression in subsequent models. The final model, which included 11 variables, explained 32% ($R^2 = .320$) of the variance in analytic word-choice scores. This model was significant, $F(11, 704)$, $p < .001$. When the model was cross-validated, it explained 30.5% of variance ($R^2 = .305$), suggesting that the model is stable across the dataset. The final model included indices related to phonological neighbors, lexical-decision times, word familiarity and frequency, and association strength. The results indicated that texts that were scored higher in word choice included more sophisticated lexical features. Table 6 presents a summary of the regression model.

Discussion

This study introduces and helps validate TAALES 2.0, an easy to use, freely available, versatile tool for measuring a wide variety of indices related to lexical sophistication. It is hoped that researchers in a variety of fields related to discourse processing, text analysis, and language assessment will find

[4] In many language analyses, nonnormality of the data is common due to a rarity of features and/or ceiling effects.

[5] See the supplementary material document entitled Analysis_Summary.xlsx (available at www.kristopherkyle.com/supplementary-materials.html) for a summary of the variables that were included and excluded, and the reasons why. This document also includes the correlation matrices for each analysis, indicating the degree of multicollinearity between variables.

Table 3 Correlations between lexical proficiency scores and the TAALES 2.0 indices

Variable	Category	*r*
COCA News Bigram Association Strength (DP)	Ngram Association Strength	.391
COCA Magazine Trigram Proportion 80k	Ngram Frequency	.381
MRC Meaningfulness AW	Psycholinguistic Norms	–.380
MRC Familiarity CW	Psycholinguistic Norms	–.360
COCA Magazine Trigram Bigram to Unigram Association Strength (DP)	Ngram Association Strength	.337
Kučera–Francis Register Range CW	Word Range	–.320
COCA News Trigram Unigram to Bigram Association Strength (DP)	Ngram Association Strength	.316
Lexical-Decision Time	Word Recognition Norms	.312
Lexical-Decision Time (standard deviation)	Word Recognition Norms	.292
LDA Age of Exposure (inverse slope)	Age of Exposure	.289
Hypernymy Nouns and Verbs (Sense Mean, Path Mean)	Semantic Network	–.288
Brysbaert Concreteness Combined AW	Psycholinguistic Norms	–.286
COCA Spoken Bigram Association Strength (MI)	Ngram Association Strength	.284
Word-Naming Response Accuracy	Word Recognition Norms	–.281
COCA Magazine Trigram Bigram to Unigram Association Strength (MI)	Ngram Association Strength	.279
COCA News Trigram Bigram to Unigram Association Strength (MI^2)	Ngram Association Strength	.277
COCA Academic Trigram Proportion 10k	Ngram Frequency	.255
Polysemy Verbs	Semantic Network	.253
Lexical-Decision Accuracy	Word Recognition Norms	–.248
COCA Fiction Trigram Bigram to Unigram Association Strength (MI)	Ngram Association Strength	.247
COCA Spoken Bigram Association Strength (MI^2)	Ngram Association Strength	.235
MRC Imageability FW	Psycholinguistic Norms	–.226
COCA Academic Frequency CW Logarithm	Word Frequency	–.226
MRC Imageability CW	Psycholinguistic Norms	–.213

DP delta P, *AW* all words, *CW* content words

TAALES 2.0 a useful mechanism to examine lexical sophistication in a variety of situations. We envision that TAALES 2.0 might prove beneficial for researchers examining the effects of text complexity on reading comprehension and text processing. Educational assessments related to language production might also be informed by the lexical features found in TAALES 2.0. Additionally, cognitive scientists might find the tool useful in helping to develop language stimuli for behavioral experiments. Computational social scientists may use the tool's features to examine trends reported in traditional or social media. Here, we examined if the indices reported in TAALES 2.0 were predictive of human judgment of lexical proficiency and word choice. We discuss these findings below.

Table 4 Summary of lexical proficiency multiple regression model

Entry	Predictors Included	*r*	R^2	R^2 Change	*B*	*SE*	β
1	COCA News Bigram Association Strength (DP)	.391	.153	.153	9.292	4.074	.117
2	COCA Magazine Trigram Proportion 80k	.496	.246	.093	5.747	0.843	.371
3	Kučera–Francis Register Range CW	.620	.385	.139	−0.355	0.113	–.222
4	Lexical-Decision Time (standard deviation)	.634	.402	.017	0.011	0.006	.090
5	LDA Age of Exposure (inverse slope)	.636	.405	.003	1.132	0.616	.102
6	Hypernymy Nouns and Verbs (Sense Mean, Path Mean)	.671	.451	.046	−0.633	0.111	–.296
7	Word-Naming Response Accuracy	.701	.491	.040	−76.169	22.968	–.154
8	COCA Magazine Trigram Bigram to Unigram Association Strength (MI)	.724	.524	.033	0.499	0.141	.157
9	Polysemy Verbs	.738	.544	.020	0.061	0.020	.139
10	COCA Academic Frequency CW Logarithm	.762	.580	.036	−1.221	0.276	–.259

Estimated constant term = 82.14, *B* = unstandardized beta, *SE* = standard error; β = standardized beta

Table 5 Correlations between word choice scores and the TAALES 2.0 indices

Variable	Category	*r*
SUBTLEXus Frequency CW Logarithm	Word Frequency	–.471
Phonological Neighbors (includes homonyms)	Word Neighbor Information	–.431
Lexical-Decision Time (*z* Score)	Word Recognition Norms	.427
Brown Frequency AW Logarithm	Word Frequency	–.418
MRC Familiarity AW	Psycholinguistic Norms	–.407
Age of Acquisition AW	Psycholinguistic Norms	.395
Brown Frequency AW	Word Frequency	–.385
COCA Fiction Trigram Bigram to Unigram Association Strength (DP)	Ngram Association Strength	.341
Word-Naming Response Time (standard deviation)	Word Recognition Norms	.329
COCA Spoken Trigram Proportion 60k	Ngram Frequency	–.300
Lexical-Decision Time (standard deviation)	Word Recognition Norms	.299
COCA Fiction Trigram Unigram to Bigram Association Strength (MI)	Ngram Association Strength	.287
COCA Fiction Bigram Association Strength (MI)	Ngram Association Strength	.283
COCA Academic Frequency AW Logarithm	Word Frequency	–.270
McDonald Co-occurrence Probability	Contextual Distinctiveness	.269
COCA Fiction Bigram Association Strength (DP)	Ngram Association Strength	.266
Orthographic Neighborhood Frequency	Word Neighbor Information	–.265
LDA Age of Exposure (inverse slope)	Age of Exposure	.255
COCA Fiction Trigram Unigram to Bigram Association Strength (DP)	Ngram Association Strength	.250
Polysemy CW	Semantic Network	–.250
BNC Spoken Bigram Frequency Logarithm	Ngram Frequency	–.246
Hypernymy Verbs (Sense 1, Path 1)	Semantic Network	.243
Free Association Stimuli Elicited	Contextual Distinctiveness	–.242
Free Association Types	Contextual Distinctiveness	.241
Polysemy Adverbs	Semantic Network	–.231
Phonological Neighborhood Frequency (homophones excluded)	Word Neighbor Information	–.220
COCA Fiction Bigram Frequency Logarithm	Ngram Frequency	–.218
COCA Academic Frequency FW	Word Frequency	.217
COCA Fiction Frequency CW	Word Frequency	–.205
Hypernymy Nouns and Verbs (Sense 1, Path 1)	Semantic Network	.195
Word-Naming Response Accuracy	Word Recognition Norms	–.195
Academic Word List All	Academic Language	.191
Brysbaert Concreteness Combined CW	Psycholinguistic Norms	–.191
COCA Academic Bigram Association Strength (AC)	Ngram Association Strength	.190
COCA Fiction Trigram Unigram to Bigram Association Strength (MI^2)	Ngram Association Strength	.187
SUBTLEXus Frequency AW	Word Frequency	–.187
COCA Academic Trigram Unigram to Bigram Association Strength (T)	Ngram Association Strength	.185
Polysemy Adjectives	Semantic Network	–.183
COCA Spoken Trigram Bigram to Unigram Association Strength (MI)	Ngram Association Strength	.183
SUBTLEXus Range FW	Word Range	–.177
COCA Spoken Trigram Range Logarithm	Ngram Range	–.170
MRC Meaningfulness AW	Psycholinguistic Norms	–.164
COCA Academic Bigram Association Strength (T)	Ngram Association Strength	.155
Character Bigram Frequency	Character Bigram Frequency	.152
COCA Academic Trigram Range	Ngram Range	.146
COCA Spoken Trigram Bigram to Unigram Association Strength (AC)	Ngram Association Strength	–.123

Table 6 Summary of word-choice multiple regression model

Entry	Predictors included	*r*	R^2	R^2 Change	*B*	*SE*	β
1	Phonological Neighbors (includes homonyms)	.431	.186	.186	−0.064	0.029	–.104
2	Lexical-Decision Time (*z* Score)	.473	.224	.038	10.030	3.119	.164
3	MRC Familiarity AW	.503	.253	.029	−0.026	0.018	–.065
4	Brown Frequency AW	.524	.274	.022	−0.001	0.001	–.069
5	COCA Fiction Bigram Association Strength (MI)	.533	.284	.010	0.718	0.372	.077
6	Orthographic Neighborhood Frequency	.539	.291	.007	−0.494	0.210	–.086
7	COCA Academic Bigram Association Strength (AC)	.542	.294	.003	0.000	0.000	.064
8	COCA Fiction Trigram Unigram to Bigram Association Strength (MI^2)	.543	.294	<.001	0.267	0.189	.058
9	COCA Academic Trigram Unigram to Bigram Association Strength (T)	.544	.296	.001	0.035	0.013	.101
10	COCA Spoken Trigram Bigram to Unigram Association Strength (AC)	.564	.318	.023	0.000	0.000	–.176
11	SUBTLEXus Frequency CW Logarithm	.566	.320	.002	−0.524	0.364	–.095

Estimated constant term = 34.08, *B* = unstandardized beta, *SE* = standard error; β = standardized beta

Lexical proficiency

Ten TAALES indices were used in a model that explained approximately 58% of the variance in lexical proficiency scores. These results are stronger than models in previous studies, which explained between 44% (Crossley, Salsbury, et al., 2011) and 51.7% (Kyle & Crossley, 2015) of the variance in lexical proficiency scores. The predictor model both supports and extends previous models of lexical proficiency. Of the ten predictor variables in the final model, only one (Kučera–Francis Register Range CW) was included in TAALES 1.4, and only two others (COCA Magazine Trigram Proportion 80k and COCA Academic Frequency CW Logarithm) are conceptually related to the TAALES 1.4 indices. This suggests that TAALES 2.0 represents an important upgrade to previous versions, both practically and conceptually.

Four indices related to *n*-gram strength of association, *n*-gram frequency, and word range contributed over two thirds of the variance explained by the model (41.8%), whereas an index related to word frequency explained only 3.2% of the variance in lexical proficiency scores. Each index category that contributed to the final model is discussed below.

N-gram association strength and frequency Indices related to *n*-gram association strength and *n*-gram frequency explained approximately 28% of the variance in lexical proficiency scores. Texts that included more strongly associated bigrams and trigrams and a higher percentage of frequent trigrams tended to earn higher lexical proficiency scores. This supports recent findings that suggest collocational knowledge is a key aspect of lexical proficiency (Bestgen & Granger, 2014; Jurafsky, Bell, Gregory, & Raymond, 2001; Kyle & Crossley, 2015; McDonald & Shillcock, 2003; Römer, 2009). The findings also suggest that *n*-gram frequency and strength-of-association indices may capture related but different aspects of collocational knowledge.

Word range One index related to word range explained 13.9% of the variance in lexical proficiency scores. Texts that included words that occur in fewer registers tended to earn higher lexical proficiency scores. This may suggest that the use of words that are more register specific is an important indicator of lexical proficiency. These results support recent findings with regard to both lexical proficiency (Kyle & Crossley, 2015) and L2 writing quality (Kyle & Crossley, 2016).

Semantic networks Two indices related to semantic networks explained 6.6% of the variance in lexical proficiency scores. Texts that included nouns and verbs with fewer hypernymic levels and that were more polysemous verbs tended to earn higher scores. This suggests that the use of less specific verbs and nouns are indicators of lexical proficiency, which supports previous findings related to lexical development (Crossley et al., 2009; Crossley et al., 2011).

Word recognition norms Indices related to word recognition norms explained 5.7% of the variance in lexical proficiency scores. Texts that included words with a wider standard deviation in lexical-decision times and that were named less accurately tended to earn higher scores. This suggests that words that are more difficult to process tend to be perceived as more sophisticated. These results generally support psycholinguistic accounts of word processing and extend psycholinguistic data to support predictions of holistic judgments of lexical proficiency in writing samples.

Word frequency One index related to word frequency explained 3.6% of the variance in lexical proficiency scores. Texts that included less frequent content words tended to earn

higher lexical proficiency scores. This negative trend aligns with previous research. The relatively limited role of frequency in the predictor model, however, suggests that other factors (e.g., *n*-gram strength of association and frequency and word range) are more directly related to the construct of lexical proficiency, supporting previous studies reporting that frequency is not the strongest predictor of word processing (Adelman et al., 2006; McDonald & Shillcock, 2001).

Age of exposure One index related to age of exposure explained a small amount of the variance (0.3%) in lexical proficiency scores. The results indicate that texts including words that have lower co-occurrence patterns at later grade level tended to earn higher scores.

Word choice

Eleven TAALES indices were used in a model that explained 32% of the variance in word-choice scores. Direct comparisons to previous studies are not possible because this is the first study that has only used the word-choice scores. However, the word-choices scores have been analyzed in combination with other scores (e.g., ideas and content, voice, and organization) using NLP features to predict overall essay scores. This analysis, which included lexical features along with other features (e.g., cohesion and syntactic complexity) explained between 40% and 52% of the variance in the overall essay scores (Shermis & Hamner, 2013). Given that we only used a single construct (lexical sophistication), the results reported here seem reasonably strong and support and extend previous models of lexical proficiency. Of the eleven predictor variables in the final model, only three (MRC Familiarity AW, Brown Frequency AW, and SUBTLEXus Frequency CW Logarithm) were included in TAALES 1.4, and these explained a relatively small portion of the variance in word-choice scores (5.3%). New variables unique to TAALES 2.0, including word neighbor information, word recognition scores, and association measures, explained the lion's share of the variance, suggesting that TAALES 2.0 represents an important upgrade to previous versions, both practically and conceptually.

Word neighbor information Two indices related to word neighbor information explained 19.3% of the variance in word-choice scores. The average number of phonological neighbors accounted for most of this variance (18.6%). Essays that included words with fewer phonological neighbors (i.e., are more phonologically distinct) tend to earn higher word-choice scores. The mean orthographic neighbor frequency score for words in a text accounted for an additional 0.7% of the variance in word-choice scores. Essays that included words with less frequent orthographic neighbors tended to earn higher scores. These results are in line with psycholinguistic findings that demonstrate that performance on naming and lexical-decision tasks is faster for low-frequency words that have more orthographic neighbors, indicating that words with fewer orthographic neighbors are more complex (Andrews, 1989; Grainger, 1990; McCann & Besner, 1987).

Word recognition norms One index related to word recognition norms explained 3.8% of the variance in word-choice scores. Essays that included words that are processed more slowly (as measured by a lexical-decision task) tended to earn higher scores. This suggests that words that are processed more slowly are considered more sophisticated by human raters. This finding is novel but is in line with previous research regarding processing difficulty (Balota et al., 2004; Forster & Chambers, 1973; Frederiksen & Kroll, 1976).

N-gram association strength Five indices related to *n*-gram association strength cumulatively explained 3.7% of the variance in word-choice scores. Essays that included bigrams and trigrams that were more strongly associated tended to earn higher scores. This finding suggests that collocational knowledge is an important aspect of lexical knowledge and is in line with previous research in L2 contexts (Bestgen & Granger, 2014) and usage-based theories regarding lexical knowledge (Römer, 2009).

Word information One index related to word information explained 2.9% of the variance in word-choice scores. Essays that included words that were less familiar tended to earn higher word-choice scores. These results align with previous studies (Crossley & McNamara, 2011; Guo et al., 2013; Kyle & Crossley, 2015).

Word frequency Two indices related to word frequency explained 2.4% of the variance in word-choice scores. Essays that included less frequent content words tended to earn higher word-choice scores. The negative relationship between corpus frequency and word-choice scores generally align with previous studies (Guo et al., 2013; Kyle & Crossley, 2015; McNamara, Crossley, & McCarthy, 2009). One word frequency index (SUBTLEXus Frequency CW Logarithm) demonstrated the strongest correlation between lexical sophistication indices and word-choice scores ($r = -.471$), underscoring the important relationship between corpus frequency and word-choice scores. The most accurate predictor model, however, weighted other indices more heavily (e.g., phonological neighbors).

Overview of findings

This study reported on two validation studies that used TAALES 2.0 indices to successfully predict holistic scores

of lexical proficiency in L2 and L1 writing (R^2 = .580), and analytic word choice scores in L1 essays (R^2 = .320). The lexical proficiency model explained greater variance in lexical proficiency scores than previous models, providing evidence to support the inclusion of the new indices included in TAALES 2.0. Furthermore, the word-choice model explained a significant amount of the variance (with a medium effect), providing further supporting evidence for the validation of the TAALES 2.0 indices.

Although correlations between TAALES 2.0 indices and word-choice scores tended to be stronger than those between TAALES 2.0 indices and lexical proficiency scores, the lexical proficiency model explained more variance than the word-choice model. One interpretation of these seemingly contradictory results is that the majority of our lexical proficiency corpus was sampled from L2 learners, which may represent a greater variation of lexical knowledge and production than found in our L1 only word-choice corpus. In the lexical proficiency model; for example, three indices related to *n*-gram association strength, *n*-gram frequency, and word range each contributed a relatively large percentage of the variance explained by the model indicating that a number of unique lexical features account for the variation in human judgments of L2 and L1 speakers. In contrast, for the word-choice model, a single index related to phonological neighbors accounted for a lion's share of the variance in scores suggesting less variation on the part of L1 writers, at least in terms of predicting word-choice scores.

The results generally support previous findings related to the lexical features reported by TAALES 1.4 (i.e., word frequency, word range, *n*-gram frequency, academic language, and psycholinguistic word information). Importantly, the results also present a number of novel findings regarding the additional lexical features included as part of TAALES 2.0 (i.e., contextual distinctiveness, word recognition norms, semantic network, *n*-gram association strength, *n*-gram range, and word neighbors). In the lexical proficiency study, for example, indices related to COCA derived *n*-gram frequency and association strength accounted for 28% of the variance in lexical proficiency scores. *N*-gram association strength indices also accounted for a portion of the variance (3.7%) in word-choice scores. Indices related to word neighbor information also accounted for over half of the variance explained by the word-choice model (19.3% of the total variance explained). Indices related to word recognition norms were included in both the lexical proficiency model and the word-choice model. Furthermore, these indices were among those that demonstrated some of the strongest correlations with lexical proficiency scores (r = .312) and word-choice scores (r = .417). Of the new TAALES 2.0 categories, the only one not represented in either predictor model was contextual distinctiveness. Although they were not included in the regression model, contextual distinctiveness indices did demonstrate significant correlations with word choice and lexical proficiency. In the case of lexical proficiency, the contextual distinctiveness indices did not meet our strict threshold of $p <$ .001 that we used to control for Type I errors.

Limitations and future directions

A major limitation of TAALES is user knowledge. The tool contains a vast repository of lexical features that most new users will not be familiar with. Thus, there will be a steep learning curve for most. Although we have endeavored to provide background information on the indices reported by TAALES in this article and in the index guide, the sheer number of indices and the lexical constructs they represent will prove daunting for the new user. In addition, statistical analyses using TAALES requires a solid knowledge of not only inferential statistics but knowledge of variable selection. For instance, experienced, but not inexperienced, users will know that frequency indices based on logarithmic transformations will more likely be normally distributed, whereas raw frequency counts may be nonnormally distributed because of Zipfian tendencies in language. Likewise, experienced users will know that some of the variables reported by TAALES depend on word vectors that are not densely populated. For instance, a number of the Academic Word Sublists contain a small number of words that are rare in smaller texts, leading to nonnormally distributed data. However, for larger texts, these indices will report normal distributions. Of course, depending on the statistical analysis, normal distributions need not be an assumption. This is especially true for a number of machine learning techniques. In addition, TAALES is purposefully redundant in that it calculates a number of lexical features for a single lexical construct (e.g., frequency) and also calculates a number of similar variables from a single database (e.g., raw and logarithmic frequency counts, range scores, and *n*-gram counts from COCA). Such redundancy allows users greater capabilities to make decisions specific to their research question, but also introduces the potential for multicollinearity and suppression effects within an analysis. Again, many new TAALES users will face a learning curve when making experimental and statistical decisions.

There are also limitations specific to this study. Foremost, we only looked at two cross-sectional datasets. One of the datasets was relatively small (n = 240), and the other had relatively low rating reliability. It would be fruitful for future studies to investigate larger datasets with more reliable human ratings. Longitudinal datasets may also allow researchers to observe how lexical sophistication develops over time, particularly in L2 or younger L1 participants. Future studies should also investigate the complexity of predictor models related to L1 and L2 lexical proficiency. In this study, we found more indices accounting for a larger percentage of the variance in the predominately L2 dataset than in the L1 dataset. Future

research should investigate whether these differences are attributable to speaker status (e.g., L1 vs. L2) or other variables, such as writing tasks and scoring rubrics. Importantly, the nature of the statistical analyses used here required the removal of a large number of indices to avoid multicollinearity. Depending on their research question(s), researchers who use TAALES 2.0 may consider employing statistical techniques that minimize multicollinearity effects, such as factor analysis. Finally, although TAALES 2.0 has been refined considerably, some limitations remain. For example, not all indices allow researchers to distinguish between values for content versus function words in a text (e.g., word recognition norms).

Conclusion

This study introduces a major update to a freely available text analysis tool, TAALES 2.0. This tool was designed to allow for efficient and replicable analysis of lexical sophistication in a variety of domains, including educational psychology, cognitive science, and artificial intelligence (among many others). The results suggest that the increased construct coverage of TAALES 2.0 positively influenced the predictive validity of models related to lexical sophistication, providing predictive validation of the indices reported by the tool. The results of both studies also suggest that the nature of lexical sophistication is multifaceted and complex. Furthermore, the results suggest that the construct of lexical sophistication is not restricted to the properties of words in isolation, but involves collocational knowledge. These findings provide new avenues for varied endeavors such as developing behavioral stimuli, automatically assessing speaking and writing proficiency, and investigating reading difficulty, among others.

References

Adelman, J. S., & Brown, G. D. A. (2007). Phonographic neighbors, not orthographic neighbors, determine word naming latencies. *Psychonomic Bulletin & Review, 14,* 455–459. doi:10.3758/BF03194088

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science, 17,* 814–823.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19,* 716–723. doi:10.1109/TAC.1974.1100705

Allen, L. K., Crossley, S. A., & McNamara, D. S. (2015). Predicting misalignment between teachers' and students' essay scores using natural language processing tools. In *International Conference on Artificial Intelligence in Education* (pp. 529–532). Berlin: Springer.

Allen, L. K., & McNamara, D. S. (2015). You are your words: Modeling students' vocabulary knowledge with natural language processing. In O. Santos, J. G. Boticario, C. Romero, M. Pechenizkiy, P. Merceron, P. Mitros, . . . M. Desmarais (Eds.), *Proceedings of the 8th International Conference on Educational Data Mining*. Madrid, Spain: International Educational Data Mining Society.

Altarriba, J., Bauer, L. M., & Benvenuto, C. (1999). Concreteness, context availability, and imageability ratings and word associations for abstract, concrete, and emotion words. *Behavior Research Methods, Instruments, & Computers, 31,* 578–602. doi:10.3758/BF03200738

Andrews, S. (1989). Frequency and neighborhood effects on lexical access: Activation or search? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 802–814. doi:10.1037/0278-7393.15.5.802

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General, 133,* 283–316. doi:10.1037/0096-3445.133.2.283

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., … & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods, 39,* 445–459. doi:10.3758/BF03193014.

Berger, C. M., Crossley, S., & Kyle, K. (2017). Using novel word context measures to predict human ratings of lexical proficiency. *Educational Technology & Society, 20,* 201–212.

Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing, 26,* 28–41. doi:10.1016/j.jslw.2014.09.004

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at … : Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25,* 371–405. doi:10.1093/applin/25.3.371

Biber, D., Conrad, S. M., Reppen, R., Byrd, P., Helt, M., Clark, V., . . . Urzua, A. (2004). *Representing language use in the university: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus*. TOEFL Monograph Series. Retrieved from www.ets.org/Media/Research/pdf/RM-04-03.pdf

BNC Consortium. (2007). The British National Corpus, version 3. BNC Consortium. Retrieved from www.natcorp.ox.ac.uk

Brown, G. D. A. (1984). A frequency count of 190,000 words in the London–Lund Corpus of English Conversation. *Behavior Research Methods, Instruments, & Computers, 16,* 502–532.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera 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,* 977–990. doi:10.3758/BRM.41.4.977

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46,* 904–911. doi:10.3758/s13428-013-0403-5

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology, 33A,* 497–505. doi:10.1080/14640748108400805

Coltheart, M., Davelaar, E., Jonasson, T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance IV* (pp. 535–555). Hillsdale, NJ: Erlbaum.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34,* 213–238.

Coxhead, A. (2011). The Academic Word List 10 years on: Research and teaching implications. *TESOL Quarterly, 45,* 355–362.

Crossley, S. A., Cai, Z., & McNamara, D. S. (2012, May). *Syntagmatic, paradigmatic, and automatic n-gram approaches to assessing essay quality.* Paper presented at the Twenty-Fifth International FLAIRS Conference, Marco Island, FL.

Crossley, S. A., Dufty, D. F., McCarthy, P. M., & McNamara, D. S. (2007). Toward a new readability: A mixed model approach. In

Proceedings of the 29th Annual Conference of the Cognitive Science Society (pp. 197–202). Austin, TX: Cognitive Science Society.

Crossley, S., Kyle, K., & Salsbury, T. (2016). A usage-based investigation of L2 lexical acquisition: The role of input and output. *Modern Language Journal, 100,* 702–715. doi:10.1111/modl.12344

Crossley, S. A., & McNamara, D. S. (2011). Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. *International Journal of Continuing Engineering Education and Life Long Learning, 21,* 170–191.

Crossley, S. A., & McNamara, D. (2013). Applications of text analysis tools for spoken response grading. *Language Learning & Technology, 17,* 171–192.

Crossley, S. A., Salsbury, T., & McNamara, D. (2009). Measuring L2 lexical growth using hypernymic relationships. *Language Learning, 59,* 307–334.

Crossley, S. A., Salsbury, T., & McNamara, D. (2010). The development of polysemy and frequency use in English second language speakers. *Language Learning, 60,* 573–605. doi:10.1111/j.1467-9922.2010.00568.x

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2012). Predicting the proficiency level of language learners using lexical indices. *Language Testing, 29,* 243–263. doi:10.1177/0265532211419331

Crossley, S. A., Salsbury, T., McNamara, D., & Jarvis, S. (2011). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing, 28,* 561–580. doi:10.1177/0265532210378031

Crossley, S. A., Weston, J. L., Sullivan, S. T. M., & McNamara, D. S. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication, 28,* 282–311.

Dascalu, M., McNamara, D. S., Crossley, S., & Trausan-Matu, S. (2016, February). *Age of exposure: A model of word learning*. Paper presented at the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, AZ.

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics, 14,* 159–190. doi:10.1075/ijcl.14.2.02dav

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing, 25,* 447–464. doi:10.1093/llc/fqq018

Evert, S. (2005). *The statistics of word cooccurrences: Words pairs and collocations (Doctoral dissertation)*. Germany: Universität Stuttgart.

Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Fisher, R. A. (1934). Two new properties of mathematical likelihood. *Proceedings of the Royal Society A*, *144*, 285–307. Retrieved from www.jstor.org/stable/2935559

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior, 12,* 627–635. doi:10.1016/S0022-5371(73)80042-8

Frederiksen, J. R., & Kroll, J. F. (1976). Spelling and sound: Approaches to the internal lexicon. *Journal of Experimental Psychology: Human Perception and Performance, 2,* 361–379. doi:10.1037/0096-1523.2.3.361

Grainger, J. (1990). Word frequency and neighborhood frequency effects in lexical decision and naming. *Journal of Memory and Language, 29,* 228–244. doi:10.1016/0749-596X(90)90074-A

Gries, S. T., Hampe, B., & Schönefeld, D. (2005). Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics, 16,* 635–676. doi:10.1515/cogl.2005.16.4.635

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing, 18,* 218–238.

Hoffman, P., Lambon Ralph, M. A., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods, 45,* 718–730. doi:10.3758/s13428-012-0278-x

Hyland, K. (2009). *Academic discourse*. New York, NY: Continuum.

Johns, B. T., & Jones, M. N. (2008). Predicting word-naming and lexical decision times from a semantic space model. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Cognitive Science Society Meeting* (pp. 279–284). Austin, TX: Cognitive Science Society.

Jung, Y., Crossley, S. A., & McNamara, D. S. (2015). Linguistic features in MELAB writing performances (Working Paper No. 2015–05), Georgia State University, Atlanta, GA.

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In J. Bybee & P. Hopper (Eds.), *Typological studies in language* (Vol. 45, pp. 229–254). Amsterdam, The Netherlands: Benjamins.

Kiss, G. R., Armstrong, C., Milroy, R., & Piper, J. (1973). An associative thesaurus of English and its computer analysis. *Computer and Literary Studies*, 153–165.

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

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods, 44,* 978–990. doi:10.3758/s13428-012-0210-4

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly, 49,* 757–786. doi:10.1002/tesq.194

Kyle, K., & Crossley, S. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing, 34,* 12–24. doi:10.1016/j.jslw.2016.10.003

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25,* 259–284. doi:10.1080/01638539809545028

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in l2 written production. *Applied Linguistics, 16,* 307–322. doi:10.1093/applin/16.3.307

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers, 28,* 203–208. doi:10.3758/BF03204766

McCann, R. S., & Besner, D. (1987). Reading pseudohomophones: Implications for models of pronunciation assembly and the locus of word-frequency effects in naming. *Journal of Experimental Psychology: Human Perception and Performance, 13,* 14–24. doi:10.1037/0096-1523.13.1.14

McDonald, S. A., & Shillcock, R. C. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech, 44,* 295–322.

McDonald, S. A., & Shillcock, R. C. (2003). Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science, 14,* 648–652.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2009). Linguistic features of writing quality. *Written Communication, 27,* 57–86.

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing, 23,* 35–59.

Nation, I. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review, 63,* 59–82.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers, 36,* 402–407. doi:10.3758/BF03195588

O'Donnell, M. B., Römer, U., & Ellis, N. C. (2013). The development of formulaic sequences in first and second language writing: Investigating effects of frequency, association, and native norm. *International Journal of Corpus Linguistics, 18,* 83–108. doi:10.1075/ijcl.18.1.07odo

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology, 76*(1, Pt. 2), 1–25. doi:10.1037/h0025327

Peereman, R., & Content, A. (1997). Orthographic and phonological neighborhoods in naming: Not all neighbors are equally influential in orthographic space. *Journal of Memory and Language, 37,* 382–410.

Reynolds, D. W. (1995). Repetition in nonnative speaker writing. *Studies in Second Language Acquisition, 17,* 185–209.

Römer, U. (2009). The inseparability of lexis and grammar: Corpus linguistic perspectives. *Annual Review of Cognitive Linguistics, 7,* 140–162. doi:10.1075/arcl.7.06rom

Shermis, M. D., & Hamner, B. (2013). Contrasting state-of-the-art automated scoring of essays. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 313–346). New York, NY: Routledge.

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics, 31,* 487–512.

Skalicky, S., Berger, C. M., Crossley, S. A., & McNamara, D. S. (2016). Linguistic features of humor in academic writing. *Advances in Language and Literary Studies, 7,* 248–259.

Skalicky, S., & Crossley, S. (2015). A statistical analysis of satirical Amazon. com product reviews. *European Journal of Humour Research, 2,* 66–85.

Spreen, O., & Schulz, R. W. (1966). Parameters of abstraction, meaningfulness, and pronunciability for 329 nouns. *Journal of Verbal Learning and Verbal Behavior, 5,* 459–468.

Svartvik, J., & Quirk, R. (1980). *A corpus of English conversation*. Lund, Sweden: Gleerup.

Thorndike, E. L., & Lorge, I. (1944). *The teacher's wordbook of 30,000 words*. New York, NY: Columbia University, Teachers College. Bureau of Publications.

Toglia, M. P., & Battig, W. F. (1978). *Handbook of semantic word norms*. Hillsdale, NJ: Erlbaum.

Yates, F. (1934). Contingency tables involving small numbers and the χ^2 test. *Journal of the Royal Statistical Society, 1*(2, Suppl), 217–235. doi:10.2307/2983604

Yates, M. (2005). Phonological neighbors speed visual word processing: Evidence from multiple tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 1385–1397. doi:10.1037/0278-7393.31.6.1385

Yates, M. (2009). Phonological neighbourhood spread facilitates lexical decisions. *Quarterly Journal of Experimental Psychology, 62,* 1304–1314.

Yates, M., Locker, L., & Simpson, G. B. (2004). The influence of phonological neighborhood on visual word perception. *Psychonomic Bulletin & Review, 11,* 452–457.