Reading and Writing (2020) 33:1591–1616
https://doi.org/10.1007/s11145-020-10024-6

Word and sentence level tests of morphological awareness in reading

Stuart E. Bernstein[1] · Jennifer L. Flipse[2] · Ying Jin[1] · Timothy N. Odegard[1,2]

Published online: 3 February 2020

Abstract

Common tests of morphological awareness measure both morphology and syntax by requiring participants to fit words and pseudowords into sentences by adding or removing affixes. We report the results of a study testing a new word level task. College students viewed transparent words (without phonological or orthographic shifts) and used a keyboard to indicate whether the items contained 1, 2, 3, or 4 morphemes. Morpheme counting accuracy was strongly and significantly correlated with sentence level tests of morphological awareness, also grouping with them in a factor analysis, suggesting that the tasks measure a similar construct. Morpheme counting accuracy was also strongly and significantly correlated with the word identification and passage comprehension measures from the WJ-IV. Crossed random-effects modeling showed that all tasks were sensitive to word frequency and vocabulary. However, different MA tasks varied in their sensitivity to the sublexical properties of words. Responses in the sentence level tasks were sensitive to word and bigram frequency while responses in the word level task was sensitive to base frequency and the number of morphemes. Our findings suggest that conscious knowledge of root words and affixes can be directly measured at the word level without a syntactic component to the task and that responses do capture variation in the ability to decompose complex words into their component morphemes.

Keywords Morphological awareness · Morphology · Word reading

✉ Stuart E. Bernstein
Stuart.Bernstein@mtsu.edu

[1] Department of Psychology, Middle Tennessee State University, Murfreesboro, TN 37132, USA

[2] Tennessee Center for the Study and Treatment of Dyslexia, Middle Tennessee State University, Murfreesboro, TN, USA

Introduction

Morphological awareness, an aspect of metalinguistic awareness, is defined as conscious knowledge of prefixes, roots, and suffixes (Levesque, Kieffer, & Deacon, 2017; Nunes & Bryant, 2006). English inflectional morphology is a closed class of eight patterns that do not change a word's part of speech or semantic category. These include tense markers for verbs (*ed*, *s*, and *ing*), plural markers for nouns (*s* and *es*), and comparative markers for adjectives (*er* and *est*). Children show early mastery of English inflectional morphology (Berko, 1958; Selby, 1972). In contrast, English derivational morphology is a larger system of affixes that do change part of speech and meaning. While knowledge of inflectional morphology is largely implicit, explicit knowledge can help in comprehension because when readers know one word in a family (e.g., *interest*) they can use morphological relationships to unpack multiple other words in that family (e.g., *interested*, *disinterested*; Nagy & Anderson, 1984).

Most research on morphological awareness has focused on children, tracking growth in knowledge of derivational morphology (Anglin, 1993; Berninger, Abbott, Nagy, & Carlisle, 2010; Carlisle & Fleming, 2003; Carlisle & Kearns, 2017; Rubin, 1988; Tyler & Nagy, 1989). Differences in morphological awareness is correlated with proficiency in word reading and comprehension in children (Cho, Gilbert, & Goodwin, 2013; Kearns, 2015; Melnychuk et al., 2013; Nagy, Berninger, Abbott, Vaughan, & Vermeulen, 2003). Recent studies have also begun to document that morphological awareness remains relevant to reading comprehension differences in college students (Wilson-Fowler & Apel, 2015), adults with low literacy (Tighe & Binder, 2015), and adults with dyslexia (Law, Wouters, & Ghesquière, 2015). Some researchers propose that relative strengths in morphological awareness may serve as a protective factor that supports the comprehension of written language in spite of word reading deficits (Haft, Myers, & Hoeft, 2016; Law et al., 2015). Collectively, these findings point to a current trend aimed at better specifying the role of morphological awareness in both typical and atypical forms of reading development and point to a potential area for intervention with struggling readers. However, progress in understanding the relationship between morphological awareness and comprehension is limited by the challenge of adapting tasks initially developed for children to measure adults. The current studies address this challenge.

Tests of morphological awareness

Defining and measuring the development of morphological awareness is complex because when we strive to do so, we are attempting to measure aspects of linguistic knowledge that include the semantic components of morphemes, orthographic and phonological representations of words, and syntactic components of language (Berninger et al., 2010; Kuo & Anderson, 2006). Interest in morphology has resulted in a proliferation of tasks designed to measure this construct. Yet, there are at least two limitations in existing measures of morphological awareness. The first and most serious is the narrow zone of effective prediction across ages, which is a problem given that derivational morphology develops across the lifespan. Tasks are commonly adapted for older readers by selecting less frequent words, but few studies have addressed the impact of all relevant variables (base word frequency, affix frequency, whole word frequency, transparency, length in morphemes) on item difficulty. The second limitation is the syntactic nature of most tasks, which is a problem when trying to measure morphological knowledge of lexical information. This could introduce method bias (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003), in that both the comprehension task and the morphological awareness task require syntactic processing.

Tests of morphological knowledge can be divided into two subtypes (Levesque et al., 2017). *Morphological decoding* or *morphological use* tasks measure how knowledge of morphology is used when the dependent variable is word reading or spelling accuracy. *Morphological awareness* (MA) tasks measure conscious knowledge of the composition of complex words. Performance on MA tasks is most commonly related to comprehension (Carlisle, 2000; Guo, Roehrig, & Williams, 2011; Levesque et al., 2017; Nagy et al., 2003; Wilson-Fowler & Apel, 2015). The role of morphological awareness is of interest in older readers because a large portion of new vocabulary encountered in texts is affixed versions of already known words (Nagy & Anderson, 1984). Consequently, performance on MA task batteries continues to account for unique variance in comprehension in adult readers (Law et al., 2015; Tighe & Binder, 2015; Wilson-Fowler & Apel, 2015). These MA task batteries are composed of some combination of the tasks described next.

Test of morphological structure

Carlisle (1988, 2000) created the most commonly used metric of morphological awareness, the *Test of Morphological Structure* (TMS). The TMS has two subtests: In the *decomposition* task participants strip the affixes from a derived word to fit the root word into a sentence (e.g., *Improvement. My teacher wants my spelling to* _________). In the *derivation* task participants add affixes to fit a root word into a sentence (e.g., *Farm. My uncle is a* _________.). Carlisle (2000) concluded that scores on the TMS measured morphological awareness because in a hierarchical regression they accounted for significant unique variance in reading comprehension that was not associated with vocabulary or word reading accuracy. Carlisle also observed that accuracy in morphological awareness for items with phonological and orthographic shifts to have been significantly correlated with word reading accuracy for items with similar shifts. This replicates other findings that phonological changes make morphological processing more difficult (Carlisle & Stone, 2005; Fowler, Liberman, & Feldman, 1995) and suggests that extracting base words was easier for children when the phonological representation of the base word was unaltered.

However, the TMS is limited in the range of morphological development that it can characterize. Carlisle (2000) observed a ceiling effect for the decomposition subtest performance for 5th graders, where accuracy was at 96% correct. Researchers commonly adapt Carlisle's (1988, 2000) decomposition and derivation tasks for older readers by adding words with lower frequencies and words with orthographic and phonological shifts that make the relationship between the base and affixed words less obvious. Performance on these revised tasks accounts for unique variance in word reading and spelling in typically developing high school seniors (Cooper, 2017), typically developing adults (Wilson-Fowler & Apel, 2015), adults learning to read (Tighe & Binder, 2015), and adults with dyslexia (Law et al., 2015). While these studies show that morphology continues to be relevant throughout the lifespan for all readers, it is unclear how and where this influence happens due to the syntactic and semantic components of the decomposition and derivation tasks. Consequently, many researchers have adapted lexical level tasks, such as the relatedness task (Mahony, 1994) and the analogy task (Nunes, Bryant, & Bindman, 1997).

The relatedness task

The relatedness task (Mahony, 1994) is based on Derwing's "comes from" task (Derwing & Baker, 1979). In Mahony's version, participants decompose affixed words to judge whether pairs of words are related (e.g., *bag—baggage, receive—reception, debt—debit*) or unrelated (e.g., *let—letter, dust—industry*). Given that items in the task do not have carrier sentences, there is not a syntactic component. Goodwin, Gilbert, Cho, and Kearns (2014) used 24 items from Mahony's original (1994) task as part of a three task composite administered to middle school students. While their overall test had high internal consistency (Cronbach's alpha = .92), the majority of items on their 70 item test were from other tasks and had carrier sentences, so it is unclear whether the relatedness task itself is a valid measure at the lexical level. Furthermore, the relatedness task has not shown strong validity for older students. Mahony (1994) observed scores for 11th grade AP students and college students to be near ceiling and not to be correlated significantly with SAT Verbal scores.

In an attempt to overcome developmental range problems with Mahony's relatedness task, Wilson-Fowler and Apel (2015) created a more difficult version by using lower frequency words and words with phonological and orthographic shifts that obscured the relation between base and target word pairs. Scores for undergraduate college students did not show a ceiling effect. However, the item response theory (IRT) analysis demonstrated that items had low discrimination parameters (less than .5) and or low difficulty parameters (less than − 3), motivating the authors to conclude that it was not a valid measure of morphological awareness in this group and the task was dropped from further analyses. Furthermore, both shift and non-shift items were among their easy and difficult items, which suggests that shifting and frequency may not be the only factors that determine difficulty in morphological awareness at the lexical level.

Tighe and Schatschneider (2015) also tested adults with a modified version of Mahony's (1994) relatedness task. They used Maag's (2007) multiple choice version of the task with lower frequency items (e.g., which is related to *noncombatant: comb*, *bat*, or *combat*). Their 29 item Morphological Skills Task (MST) did not show a ceiling effect for adults, and it had a Cronbach's $\alpha = .75$. Moreover, MST scores were significantly correlated with other measures of morphology that contained sentence frames/contextual cues, and a confirmatory factor analysis demonstrated the MST to group with these other measures, suggesting that all of these tasks tap the same knowledge. However, the authors acknowledged that this task could have grouped with the other tasks due to similarities in response formats across tasks in their battery. Furthermore, this task is poorly suited to a more comprehensive measurement of morphology because targets would be limited to words for which there are embedded unrelated words (e.g., *noncombatant* only works because *comb* and *bat* are embedded and unrelated to the word's meaning). The task is also nonspecific to morphologically processing the target items since the frequency and familiarity of the non-target items will also contribute variance in responses.

The analogy task

The analogy task ("*push : pushed :: lose:* ______"*)* developed by Nunes and colleagues (Nunes et al., 1997) also does not use carrier sentences and it likely draws on lexical level representations of base words and knowledge of affixation without tapping syntactic knowledge. Performance on this task was significantly related to the development of correctly spelling the past tense of words for children between the ages of 6 and 10. To extend the use of this task to an older population, Tighe and Schatschneider (2015) tested adult basic education students with a 15-item version of the analogy task. The task showed strong reliability with a Cronbach's $\alpha = .81$. Their analyses also demonstrated performance of the task to be strongly related to other morphology tasks. However, they did not study how this task was related to reading outcomes, nor did they examine whether or not items remained reliable when testing older readers with higher abilities.

Dimensionality of morphological awareness

Understanding the nature of morphological representations is complicated by the variety of MA tasks. Tighe and Schatschneider (2015) compared scores on seven morphological awareness tasks and two vocabulary measures completed by adult basic education students. In their confirmatory factor analysis there was a distinction between real word and pseudoword tasks but inflectional and derived morphology were not separate. Tasks with and without sentence frames also loaded on the same factors, suggesting these variants of morphological awareness tasks measure the same aspect of morphology. Similarly, Spencer et al. (2015) found that a single factor model was the best fit for nine tests of morphological awareness and two tests of vocabulary completed by fourth graders. However, these outcomes contradict one from Goodwin, Petscher, Carlisle, and Mitchell (2017), who tested seventh and eighth graders with a battery of morphological knowledge, vocabulary, and reading comprehension tasks. Their seven morphological knowledge tasks did not fit a smaller number of dimensions in a confirmatory factor analysis. Nevertheless, Goodwin et al. (2017) suggest a resolution in which a general factor represents understanding of a word's morphological structure and a second component reflects the application of morphological knowledge to perform other aspects of literacy.

Dimensionality of lexical representations

Understanding the dimensionality of MA tasks is of interest as a tool to reveal the dimensionality of the lexicon as it pertains to the systems that support the comprehension of written language (Perfetti, 2007; Perfetti & Stafura, 2014). It is through an understanding of these factors that we can further advance our knowledge of the structure and function of interrelated constructs that support reading development across the lifespan. Factors that are relevant to performance in MA tasks (base word frequency, word frequency, and the presence of phonological and orthographic shifts) is a subset of factors that have been studied at the item level in word recognition. Word recognition is influenced both by lexical factors of word frequency and vocabulary size and sublexical factors such as neighborhood size, bigram frequency, base frequency, word family properties, and affix frequency (Chetail, Balota, Treiman, & Content, 2015; Goodwin et al., 2014; Kearns, 2015; Kieffer, Petscher, Proctor, & Silverman, 2016; Ulicheva, Harvey, Aronoff, & Rastle, 2018; Yap & Balota, 2009). Many theories of the lexicon include morphological representations, such that words are represented as both whole items and as component parts (for a discussion see Chetail et al., 2015).

Perfetti's lexical quality hypothesis (Perfetti, 2007; Perfetti & Hart, 2002; Perfetti & Stafura, 2014) explains how morphology is related to other lexical properties in reading. From this perspective, efficient reading at the item level is supported by four knowledge components: orthography, phonology, grammar (morpho-syntactic inflections), and meaning. During development, these representations become more fully specified, and with enough exposure these constituents become bound, resulting in stable representations and the synchronous delivery of information from each system. Thus, with experience, representations become lexical amalgams, rich and complex representations of multiple components (Chetail et al., 2015; Ehri & Robbins, 1992).

Current experiment/hypotheses

Research on morphological awareness with adults has focused on properties of the tasks (e.g., Tighe & Schatschneider, 2015) and the statistical potential for items to discriminate (e.g., Wilson-Fowler & Apel, 2015). Paradoxically, while studies of adults do show that differences in comprehension are significantly related to performance on MA task batteries, studies also have shown that sentence level (Carlisle, 2000) and word level (Wilson-Fowler & Apel, 2015) tasks lose their ability to discriminate among older readers. In order to advance understanding of MA in adults, a comprehensive study of the factors governing item difficulty is needed. Studies of item difficulty need a broad test with enough items to capture variance at multiple levels—which would involve an impractically long test with purely sentence level items. A lexical level test of morphological awareness, valid for adults, would be a useful part of such a battery in order to expand the item set for analysis.

To address the need to measure morphological awareness at the lexical level we chose to develop a morpheme counting task that follows the logic of syllable and phoneme counting tasks. Awareness of syllables and phonemes has been defined as the ability to count them and counting accuracy of syllables develops before the counting accuracy of phonemes (Liberman, Shankweiler, Fischer, & Carter, 1974). Treiman and Baron (1981) found that for first graders, phoneme counting was strongly and significantly correlated with nonsense word reading and syllable counting was strongly and significantly correlated with scores on a comprehension test. Nunes and Bryant (2006) suggested that counting morphemes reflects explicit awareness of them. However, no study has explored whether morpheme counting is a valid means of assessing morphological awareness. To address this need, the current study was conducted as a first step in developing and characterizing a morpheme counting task in older readers. Our primary objective was to assess the construct validity of morpheme counting by measuring how it is related to other established tests of morphological awareness. Our secondary objective was to explore the factors that govern the difficulty of items on morphological awareness tests. A better understanding of what makes a polymorphemic word challenging is necessary for the development of reliable and valid measures of morphological awareness that can be used across the lifespan and different populations of readers. It is also important for better informing our understanding of morphology.

We used item response theory to select a set of words for morpheme counting that had item difficulties within a desired range and then addressed three research questions with the newly created task. First, we examined concurrent validity by measuring the extent to which scores in morpheme counting were related with other established measures of morphological awareness in correlations, regressions, and exploratory factor analysis. Second, we explored the construct validity of word and sentence level MA tasks in the same exploratory factor analysis. Our logic was that tasks that measure the same constructs should be similarly associated with variables that measure vocabulary, decoding, and word reading. Third, we tested the assumption that overall performance in the battery of MA tasks would show sensitivity to factors that indicate morphological analysis (i.e., base word frequency, bigram frequency, and the number of morphemes).

For this final question, we used crossed random models to address how item level factors (shift, affix family frequency, bigram frequency, affix frequency, word frequency), participant level factors (vocabulary, word reading accuracy, and decoding), and task (sentence vs word) were related to response accuracy in morphological awareness tasks. As has been discussed by others (Baayen, Davidson, & Bates, 2008; Carson & Beeson, 2013), the use of this statistical procedure includes both item level and person level factors within the same model, overcoming the challenges of the quasi F approach traditionally adopted in psycholinguistic studies (Clark, 1973).

Methods

Participants

Undergraduate college students ($n = 114$) participated for course credit. All were native speakers of English and there were no other exclusionary factors. As such, scores reflected the full range of ability seen in the population of undergraduates at a state university—standard scores on WRMT subtests and the TOWRE ranged from 60 to 130. Age ranged from 18 to 45 ($M = 21$, $SD = 4$). Average standard scores for tests in the battery are presented at the bottom of Table 1.

Materials

Woodcock Reading Mastery Tests 3rd Edition (WRMT-III; Woodcock, 2011)

The WRMT is an individually administered standardized test battery for word and subword reading skills. The following subtests were used and reliability coefficients were calculated for the respondents in our sample. Phonological decoding was assessed with the word attack subtest ($\alpha = .800$). Single word reading accuracy was assessed using the word identification subtest ($\alpha = .799$). Vocabulary was assessed with the word comprehension subtest ($\alpha = .897$). Comprehension of written text passages was assessed with the passage comprehension subtest ($\alpha = .815$).

Test of Word Reading Efficiency 2nd Edition (TOWRE-2; Torgesen, Wagner, & Rashotte, 2012)

Word reading efficiency was measured using Sight Word Efficiency subtest from the Test of Word Reading Efficiency-Second Edition that requires a child to pronounce as many words as possible in 45 s. Phonological decoding efficiency was measured using the Phonological Decoding Efficiency subtest that required participants to pronounced as many pseudowords as possible in 45 s.

Table 1 Item characteristics for the morphological awareness tasks

Characteristic	Task		
	MTMA	MCT—core	MCT—expanded
Log word frequency	7.18 (2.39)	5.94 (2.35)	6.24 (2.01)
Log base frequency	8.91 (1.80)	10.49 (1.87)	9.69 (2.03)
Log bigram frequency	3.63 (0.08)	3.64 (0.17)	3.58 (0.20)
Average log family frequency	6.68 (1.62)	6.18 (1.32)	6.45 (1.21)
Number of morphemes	2.43 (0.74)	3.00 (0.83)	2.80 (0.76)
Number of items in test	30	24	40

MTMA Modified Test of Morphological Awareness (Cooper et al., 2015), *MCT* Morpheme Counting Task, core set of 24 items post-IRT selection. *MCT—expanded* Morpheme Counting Task with expanded item set

Morpheme counting task (MCT)

Initial materials for this test were 60 words, 15 each with one, two, three, and four morphemes. Lexical properties were retrieved from the English Lexicon Project site (Balota et al., 2007). We only counted morphemes that are productively used in English. For instance, *radioactive* is listed as <radio<{act}>ive>=3 morphemes, but *conducted* is listed as {con—duct}>ed>. We avoided words like *conducted* because their Latin roots are not productively used in English. Words in the categories were equated for the number of phonemes and the number of syllables so that these lengths did not provide clues to the correct answers for the morpheme counting response. Items with phonological and/or orthographic shifts were minimized in this initial study of the task (there were only four shift items: *renamed*, *traders*, *homecomings*, and *removals*) because they could be processed holistically and might reduce the sensitivity of the test to segmentation.

There was an initial block of 12 practice trials and participants continued practice with the same 12 items until they answered 8 items correct in a row. Practice was followed by 56 experimental trials. An initial IRT analysis of responses from the first 70 participants was used to reduce this to a core set of 32 items: 8 each of words with 1, 2, 3, and 4 morphemes. While there were no items beyond the ± 2 logit cutoffs, there was redundancy in item difficulty, indicating that the test was longer than necessary. Some items with redundant difficulty values were trimmed from the test with the constraint that an equal number of items with each number of morphemes were retained. The final set of items appears in "Appendix 1". Following this reduction, an additional trial block was added to the end of MCT containing 16 items with phonological and orthographic shifts. There were 8 each of the two and three morpheme length and they appear in "Appendix 2". These items were only presented to the final 44 participants.

Modified Test of Morphological Awareness

This test was used to measure morphological awareness. The Modified Test of Morphological Awareness (MTMA) is a 30 item, group administered measure developed from Carlisle's original Test of Morphological Awareness (Carlisle, 2000) for use with older students. It measured both derivational and decomposition processes of morphological awareness. Five items from each process of Carlisle's task were included as the initial trials to provide appropriate floor of the test followed by additional items of lesser frequency included to avoid ceiling effects originally detected with young children. Reliability coefficient in the current sample of college students, $\alpha = .85$, is similar to published ranges of .83–.85 for a sample of high school students (Cooper, Coggins, & Elleman, 2015).

Morphological nonword analysis task

This is a measure of morphological analysis that underlies the process of morphological generalization. This 18 item multiple-choice measure contains nine morphologically accessible and nine morphologically inaccessible nonwords. Each of the nonwords was presented in a context sentence followed by three answer choices. The choices were comprised of answers ranging from single word to short phrase in length. Each of the answer choices was semantically plausible in the sentence. For example, the context sentence for the nonword *addicant* is: The *addicant* was removed from the store shelves; the possible answer choices are: *problematic drug, expired food*, and *fire-causing chemicals*. To determine the intended meaning of the nonword, addicant, the participant must analyze the morphological composition of the word to determine the correct answer is problematic drug. Reliability for the current sample of college students, α = .78, is similar to the published value from McCutchen & Logan, (2011) for this measure, α = .74.

Lexical and sublexical statistics

Item characteristics for the word and sentence level MA tasks are presented in Table 1. Frequencies for words, base words, and bigrams were taken from the English Lexicon Project database (Balota et al., 2007). Average Family Frequency (AFF) was calculated by first extracting word families from CATVAR 2.0 (Habash & Dorr, 2003) then looking up each item in the English Lexicon Project database and keeping only items that matched word bodies with the initial target as specified in the morphological spellings. Average frequency of the matched items was then calculated as AFF. Words with two bodies were represented as the average of a search for each item separately.

Procedure

Participants were tested in pairs. They first completed a task from another study not reported in this manuscript, then in a balanced order, one participant went into an adjacent room with the examiner and completed the WRMT and TOWRE while a second participant completed the morpheme counting task and other morphological awareness tasks on a computer. All tasks in the MA battery were administered by computer using PsyToolkit (Stoet, 2010) with printed stimuli and typed responses for ease and consistency of administration.

The morpheme counting task was administered by computer and was programmed in PsyToolkit (Stoet, 2010). Participants were instructed that morphemes are parts of words that have meaning; these include prefixes, root words, and suffixes. They were told to use the keyboard to count the number of morphemes in words and given the examples that words like suppose have one morpheme, words like training = train + ing have two, and words like retraining = re + train + ing have three. Reaction time to press a number key was collected. Words appeared individually in a bold sans-serif font in the middle of the screen. Correct responses were followed by a feedback screen repeating the correct response, such as "correct, walking has two morphemes, walk + ing." Incorrect responses were also followed by a feedback screen, such as "incorrect, walking has two morphemes, walk + ing." Responses that took longer than a 6 s deadline timed out and were followed by a warning to respond faster.

The task began with an initial block of 12 practice trials and participants continued practice with the same 12 items until they answered 8 items correct in a row. Practice was followed by 56 experimental trials. Feedback after errors and correct responses continued through the experiment trials but trials with errors were not repeated.

Results

Means and standard deviations by participants for all scores in the morphology test battery appear in Table 2 and average accuracy for each item in the morphology test are presented in the "Appendices 1 and 2".

IRT analysis for the morpheme counting task

The Rasch model within the IRT framework was used to analyze the psychometric properties of items in MCT. Under the Rasch model, item discrimination parameters are fixed to 1 and item difficulty parameters are estimated in the scale of logits (e.g., items with smaller logit scores are easier items). Linacre (1994) explained that with sample size around 100, consistent parameter estimates (i.e., within ± 0.5 logit) can be obtained using the Rasch model. Item difficulty parameters and their standard errors appear in the "Appendices 1 and 2". Figure 1 shows the person-item map, in which the histogram describes the distribution of students' estimated ability in MCT and the dots represent difficulty parameters of items in MCT. The person-item map indicates that these items can measure a wide range of students' ability in MCT, with most items targeting students whose ability in MCT is in between 0 and − 2 logits. The test information function presented in Fig. 2 further confirms that when students' ability in MCT is between 0 and − 2 logits, their true ability is most precisely estimated by these items (i.e., standard errors are lowest between 0 and − 2 logits).

Correlations, regression, and factor analysis

Correlations of tasks in the battery appear in Table 2. Accuracy in all the MA tasks, including MCT, were significantly related to reading outcomes. For morpheme counting, reaction time was not significantly related to any outcomes. Consequently, overall accuracy in morpheme counting was used in all subsequent analyses. It is also worth noting that the morpheme counting task reduced to 32 items (8 per category) post-IRT selection, and it was correlated with word comprehension and passage comprehension as strongly as the derivation and decomposition tasks.

A set of hierarchical regressions were conducted to explore the relationship between MCT and reading outcomes. Our question was whether how much unique variance MCT scores would predict in the second step of a hierarchical regression, after variance associated with other factors was entered. These appear in Table 3, where it can be seen that MCT scores account for unique variance in

Table 2 Descriptive statistics and bivariate correlations for all tasks

	1	2	3	4	5	6	7	8	9	10	11	12	13
1. WRMT listen. comp.	–	.521**	.610**	.452**	.226*	.161	.195*	− .026	.440**	.460**	.453**	.359**	.550**
2. WRMT passage comp.		–	.694**	.618**	.347**	.200*	.349**	.175	.519**	.527**	.499**	.429**	.617**
3. WRMT word comp.			–	.704**	.397**	.292**	.342**	.124	.590**	.596**	.655**	.493**	.740**
4. WRMT word ID				–	.726**	.465**	.679**	.217*	.398**	.585**	.477**	.370**	.553**
5. WRMT word attack					–	.397**	.657**	.191*	.244**	.328**	.324**	.246**	.349**
6. TOWRE PDE						–	.606**	.014	− .084	.254**	.159	.090	.101
7. TOWRE SWE							–	.085	.060	.296**	.158	.137	.164
8. MA: MCT RT								–	.211*	.111	.107	.146	.164
9. MA: MCT accuracy									–	.443**	.504**	.446**	.870**
10. MA: derivation										–	.624**	.492**	.718**
11. MA: decomposition											–	.464**	.780**
12. MA: nonword choice												–	.677**
13. MA: all tasks													–
M	93	99	95	96	92	100	99	3139	61	55	76	65	63
SD	13	13	13	13	16	11	12	670	29	14	18	15	13

WRMT Woodcock Reading Mastery Tests, 3rd edition (Woodcock, 2011); *word ID* word identification; *TOWRE* Test of Word Reading Efficiency, 2nd edition (Torgesen et al., 2012); *PDE* Phonemic Decoding Efficiency; *SWE* Sight Word Efficiency; *MA* Morphological Awareness; *MCT* Morpheme Counting Task. Scores on the WRMT and TOWRE are standard scores, based on adult norms. Scores on MA tasks are percent of items correct

*$p < .05$; **$p < .01$; ***$p < .001$

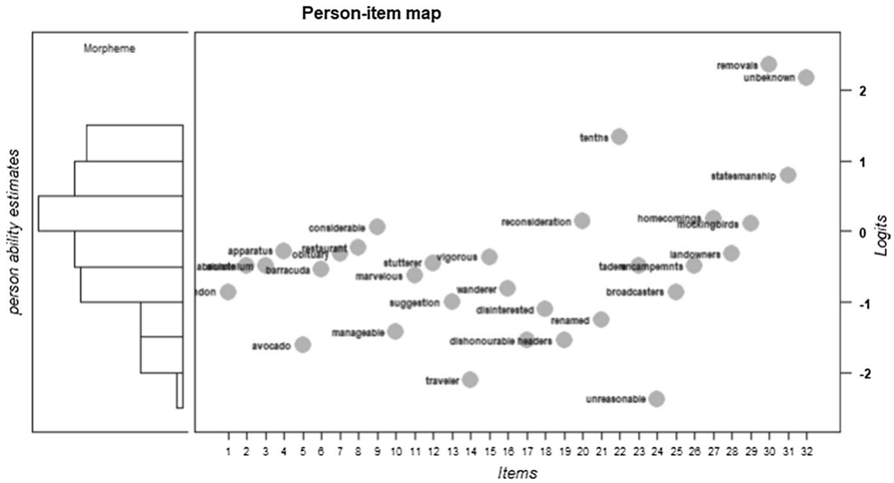

Fig. 1 Person-item map for MCT

word comprehension and passage comprehension from the WRMT but not in word identification.

An exploratory factor analysis was conducted to evaluate the concurrent validity of morpheme counting. Our logic was that a valid task should associate with other established MA tasks. We also tested the hypothesis that tasks measuring the same underlying construct should also have similar associations with tests of other word reading skills (i.e., word attack, word ID, vocabulary). Ten tasks in the battery were factor analyzed using principal component analysis with Oblimin (oblique) rotation. The analysis yielded two factors explaining a total of 65% of the variance for the entire set of variables. The rotated pattern matrix is shown in Table 4. All of the MA tasks loaded together in a single factor, labeled morpho-semantic because it also includes listening comprehension, passage comprehension, and word comprehension. This first factor explained 47% of the variance. The second factor derived was labeled decoding due to the high loadings by word attack, word ID, and both subtests of the TOWRE. None of the MA tasks loaded onto this second factor. The variance explained by this factor was 18%.

Crossed random-effects models

Items from the three tests of morphological awareness with word targets (derivation, decomposition, and morpheme counting) were combined to form a single large dataset of 48 items. The purpose of the analyses was to determine how the tasks differ in their sensitivity to sublexical properties (base frequency, bigram frequency, and the number of morphemes) and whether these item properties interact with frequency, words with lower frequencies would show greater sensitivity to sublexical properties. Item characteristics (word frequency, base word frequency, and bigram frequency) were taken from the English Lexicon Project database (ELP; Balota et al., 2007). Average family frequency (AFF), the average frequency of categorical

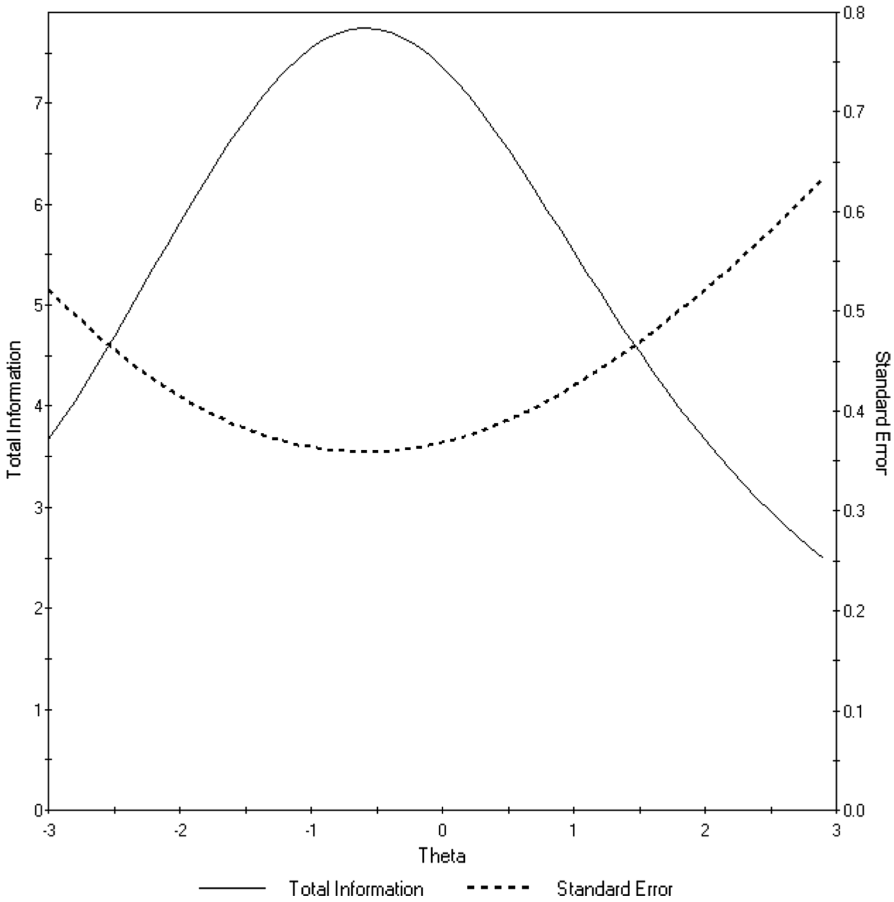

Fig. 2 Test information function for MCT

Table 3 Hierarchical regression examining the relationship between MCT and reading outcomes

Step	Factor	Word ID		Word comp		Passage comp	
		β	Δr^2	β	Δr^2	β	Δr^2
1	WRMT word attack	.530***	.733***	− .242*	.523***	− .102	.518***
	WRMT word ID	–		.880***		.342**	
	WRMT word comp	.188***		–		.492***	
2	MCT	− .021	.001	.357***	.107***	.182*	.022*

WRMT Woodcock Reading Mastery Tests, 3rd edition (Woodcock, 2011), *word ID* word identification, *word comp* word comprehension, *Passage comp* passage comprehension, *MCT* Morpheme Counting Task Accuracy

Table 4 Results of an exploratory factor analysis with all tasks in the test battery

	Component	
	1	2
Pattern matrix[a]		
WRMT word comprehension	.812	.143
MA: decomposition	.775	−.048
MA: MCT	.735	−.154
MA: derivation	.652	.163
WRMT listening comprehension	.637	.021
MA: nonword choice	.588	.024
TOWRE PDE	−.121	.953
WRMT word attack	.156	.679
WRMT word ID	.439	.658
TOWRE SWE	−.042	.645

Extraction method: principal axis factoring

Rotation method: Oblimin with Kaiser normalization

WRMT Woodcock Reading Mastery Tests, 3rd edition (Woodcock, 2011), *word ID* word identification, *TOWRE* Test of Word Reading Efficiency, 2nd edition (Torgesen et al., 2012), *PDE* Phonemic Decoding Efficiency, *SWE* Sight Word Efficiency, *MA* Morphological Awareness, *MCT* Morpheme Counting Task

[a]Rotation converged in 5 iterations

variants of the words (e.g., suggestion: suggest, suggested, suggestive, suggestible, suggestively, suggestibility) was calculated by first generating the categorical variants of the words using catvar 2.0 (Habash & Dorr, 2003) and then retrieving frequencies from the ELP database. Items without entries in ELP ($n=3$) and 1 morpheme items from MCT ($n=8$) were excluded from this analysis.

Item level accuracy was modeled using a crossed random-effects model implemented using the lme4 package (Bates, Mächler, Bolker, & Walker, 2018) within R. Model fitting began with a base model, model 0, in which the probability of a participant correctly responding to an item is the grand mean across persons plus a person-specific and an item-specific random effect.

The base model for the combined tasks had a logit intercept of 0.4021, indicating that participants had a 60% probability of correctly responding to an item. Variance observed in the outcome variable was partitioned into random effects and variance explained by participants and items. Next we ran three additional models, which appear in Table 4, each of which included a random intercept for item and participant. Model 1 evaluated how the grand mean centered item level properties common to all tasks (log word frequency, log affix frequency, log base word frequency, log bigram frequency, and log average family frequency) varied across tasks (derivation, decomposition, and MCT) and the person level properties (word attack, word ID, and vocab/word comp).

In the overall model, MA task accuracy was positively related to word frequency and vocabulary. There were no significant effects of sublexical item properties (base word frequency, bigram frequency, number of morphemes). Frequency significantly interacted with vocabulary/word comp, but critically word frequency did not interact with base word frequency, bigram frequency, or number of morphemes. That is, there was no tendency for greater sensitivity to sublexical factors for lower frequency items. Test type did interact significantly with some item and person level factors—this was subsequently explored in the next two models in Table 5, where tasks were modeled separately.

The interactions of test with items and persons were explored by analyzing responses separately for the sentence tasks and MCT. In both models, vocabulary/word comprehension and frequency acted as they did in the overall model. The only difference between the two models was the large and significant effect of bigram frequency in the sentence level tasks. There were no effects of any other sublexical factors, most notably no effect of base word frequency. Importantly, in the separate task models, as in the overall models, there was also no interaction of whole word and base word frequency.

A final pair of models is presented in Table 6. In these models, the final 44 participants in the MCT task were given additional trial blocks with shift items. We ran two models. The core model was a repetition of the MCT model from Table 5, with the original 24 items, but restricted to the 44 participants in the shift model. The shift model is for an MCT item set with additional 8 items each in the two and three morpheme lengths that had phonological and orthographic shifts. Some core items also classified as shift items are marked in the "Appendices 1 and 2". The core item model with 44 participants replicates the MCT model with 114 participants. In the shift model, even though shifting did not influence accuracy, in the expanded item set, MCT accuracy was significantly impacted by two item level sublexical properties: base word frequency and the number of morphemes.

Discussion

This experiment was motivated by the need for a word level morphological awareness task that is valid for older readers. We created a word level task, morpheme counting, and assessed its validity though comparisons to other established tests of morphological awareness and other reading related tasks. The sentence and word level MA tasks were correlated and shared similar associations with other tasks in the battery, suggesting they measure similar things. However, crossed random effects modeling showed that different MA tasks varied in their sensitivity to the sublexical properties of words.

First research question: concurrent validity

One goal of the current study was to create a task that measures morphological awareness at the lexical level. The morpheme counting task does not have syntactic demands, which is relevant because Guo et al. (2011) demonstrated that MA and syntax are separate sources of variance in comprehension for adults. Our initial

Table 5 Crossed random effects models predicting morphological awareness task accuracy

Fixed effects	Combined tasks			Sentence tasks			MCT		
	Est.	SE	z	Est.	SE	z	Est.	SE	z
Intercept (γ_{000})	1.382	1.019	1.356	– 0.303	1.203	– 0.252	0.177	0.993	0.179
Participant covariates									
γ_{001} word attack	0.013	0.025	0.529	– 0.008	0.039	– 0.196	0.029	0.031	0.930
γ_{002} word ID	0.026	0.034	0.753	0.046	0.052	0.879	– 0.012	0.042	– 0.295
γ_{003} word comp	0.093	0.013***	6.930	0.083	0.017***	4.964	0.054	0.012***	2.519
Item covariates									
γ_{010} word freq.	0.403	0.183*	2.200	0.756	0.208***	3.632	0.281	0.112*	2.519
γ_{020} base freq.	– 0.077	0.181	– 0.424	– 0.241	0.189	– 1.275	– 0.085	0.147	– 0.581
γ_{030} bigram freq.	1.278	1.631	0.783	11.041	4.019**	2.748	0.576	1.159	0.497
γ_{040} AFF	0.295	0.153	1.920	0.380	0.310	1.229	– 0.040	0.133	– 0.302
γ_{050} nmorph	0.110	0.316	0.347	0.437	0.443	0.986	0.093	0.325	0.286
γ_{060} derivation	– 2.128	0.604***	– 3.525						
γ_{070} MCT	– 0.825	0.507	– 1.628						
Item × item interactions									
γ_{101} word freq × base freq.	0.065	0.053	1.233	0.022	0.139	0.159	0.059	0.042	1.438
Item × Person interactions									
γ_{201} word comp × word freq.	0.007	0.003*	2.537	0.005	0.005	0.992	0.007	0.003*	2.301
γ_{202} word comp × base freq.	– 0.003	0.003	– 0.954	0.002	0.004	0.375	– 0.003	0.003	– 0.962
Test × item interactions									
γ_{301} derivation * word freq.	0.500	0.228*	2.195						
γ_{302} MCT * word freq.	– 0.088	0.220	– 0.400						
γ_{303} derivation * base freq.	– 0.246	0.287	– 0.856						
γ_{304} MCT * base freq.	0.038	0.216	0.177						

Table 5 (continued)

Fixed effects	Combined tasks			Sentence tasks			MCT		
	Est.	SE	z	Est.	SE	z	Est.	SE	z
Test × Person interactions									
γ_{401} MCT * word comp.	− 0.035	0.015*	− 2.350						
γ_{402} derivation * word comp.	− 0.037	0.013**	− 2.935						
Random effects	SD	Corr.	Reduc.	SD	Corr.	Reduc.	SD	Corr.	Reduc.
r_{001} participant	0.515		0.454	0.795		0.748	0.658		0.262
r_{011} wfreq	0.051	0.082		0.051	1.000		0.038	1.000	
r_{002} item	1.130		2.590	1.356		4.670	0.884		0.521
r_{012} word comp	0.047	0.072		0.031	− 0.170		0.015	1.000	
r_{022} word ID	0.005	0.097		0.053	− 0.370		0.037	0.200	

Word Attack, *Word ID* word identification, *word comp.* word comprehension—all subtests of the Woodcock Reading Mastery Tests, 3rd edition (Woodcock, 2011), *AFF* Average Family Frequency, *Nmorph* number of morphemes, *MCT* morpheme counting test, *Corr.* correlation between random effects, *Reduc.* reduction in variance from unconditional model

$^*p < .05$; $^{**}p < .01$; $^{***}p < .001$

Table 6 Crossed random effects models comparing two versions of the morpheme counting task

Fixed effects	Core items			Expanded shift items		
	Est.	SE	z	Est.	SE	z
Intercept (γ_{000})	1.321	1.268	1.042	1.879	0.681**	2.756
Participant covariates						
γ_{001} WRMT word attack	0.115	0.076	1.528	0.105	0.058	1.813
γ_{002} WRMT word ID	– 0.110	0.088	– 1.257	– 0.093	0.068	– 1.358
γ_{003} WRMT word comp	0.087	0.026***	3.293	0.061	0.021**	2.960
Item covariates						
γ_{010} word frequency	0.217	0.129	1.680	0.066	0.087	0.762
γ_{020} base frequency	– 0.010	0.157	– 0.062	0.166	0.082*	2.002
γ_{030} bigram frequency	2.120	1.345	1.576	1.162	0.654	1.775
γ_{040} AFF	0.012	0.174	0.067	0.010	0.118	0.086
γ_{050} nmorph	– 0.291	0.414	– 0.703	– 0.600	0.238**	– 2.514
γ_{060} shift				– 0.081	0.289	0.282
Item × participant interactions						
γ_{101} word comp. * word freq.	0.006	0.004	1.332	0.001	0.004	0.191
γ_{102} word comp. × base freq.	0.004	0.005	0.740	0.007	0.004	1.844
Random effects	SD	Corr.	Reduc.	SD	Corr.	Reduc.
r_{001} participant	0.878		0.403	0.439		0.183
r_{011} word freq.	0.056	1.000		0.000	−1.000	
r_{002} item	0.970		0.656	0.694		0.282
r_{021} word comp.	0.019	1.000		0.001	0.980	
r_{022} word ID	0.016			0.002	−0.390	

Word Attack, *Word ID* word identification, *word comp.* word comprehension—all subtests of the Woodcock Reading Mastery Tests, 3rd edition (Woodcock, 2011), *AFF* Average Family Frequency, *Nmorph* number of morphemes, *MCT* morpheme counting test, *Corr.* correlation between random effects, *Reduc.* reduction in variance from unconditional model

*$p<.05$; **$p<.01$; ***$p<.001$

tests of this task were done with college students and focused on validating the task. The test was administered with a set of 56 items and was reduced to a core of 32 items that showed acceptable difficulty values in IRT, within 2 logits from 0, and did not have redundant item difficulty. Elimination of items was also done to keep a rough equivalence in item length so that it would not serve as a cue to the number of morphemes. Values for difficulty were at the lower end of the ideal range, falling between 0 and – 2 logits, which would be expected given the sample of college students.

Comparisons across tasks helps establish the construct validity of morpheme counting. Accuracy was strongly and significantly correlated with other assessments of morphological awareness: decomposition, derivation (Carlisle, 2000; Cooper et al., 2015), and nonword choice (McCutchen & Logan, 2011). In the hierarchical regressions, morpheme counting accounted for unique variance in passage comprehension and word comprehension but not word identification, suggesting that the task is more sensitive to processing of meaning rather than decoding. This interpretation is consistent with the outcome for reaction times in MCT, which were relatively long and not correlated with other tasks. The RT findings suggest that the task reflects morphological problem solving and aspects of working memory rather than lexical access—this bears further exploration.

The exploratory factor analysis showed that the word level task, morpheme counting, grouped with the sentence level tasks, decomposition and derivation, replicating the association found with confirmatory factor analyses (Spencer et al., 2015; Tighe & Schatschneider, 2015). The correlations and factor analyses suggest that the morphological awareness tasks measure a single underlying construct. The factor analysis also showed that average performance on each of the MA tasks grouped much more strongly with word meaning related variables (i.e., word ID and listening comprehension) than decoding related variables (word attack, SWE and PDE), which raises concerns about the construct validity of all the MA tasks as they are typically administered with relatively short lists of items—do they measure sublexical processing of morphology?

Construct validity: sublexical processing

We used crossed random effects modeling to explore item level differences in lexical, sublexical, task, and person components. The hypothesis we tested was that decomposition into morphemes would be indicated when item accuracy in MA tasks was positively related to base word frequency, bigram frequency, and length in morphemes. With the exception of an effect of bigram frequency in the sentence tasks, this was not true of any of the initial analyses. Furthermore, base word frequency did not interact with whole word frequency in any the three initial analyses, suggesting that adjusting frequency down would not increase morphological analysis in these tasks. The first three CREM analyses confirmed what was also seen in the exploratory factor analysis, the items and tasks we used were not sensitive to individual differences in the ability to divide complex words into their component morphemes.

Having established the concurrent validity of MCT, it performs like other MA tasks, we next used the task as intended—adding blocks of more difficult items to the end of the task, allowing us to both keep the core item set intact for the analyses in Table 5, and explore the effects of shift items in Table 6. Unlike the sentence level tasks, the expanded item set MCT task showed sensitivity to base frequency and the number of morphemes. There is a vocabulary component, but it is relatively small compared to the sublexical influences: higher accuracy for more frequent base words and lower accuracy for longer words. This accomplished what we intended in the task design—understanding when sublexical factors matter in MA tasks. Both the failures in Table 5 and the successes in Table 6 are relevant, since most of us build our task batteries with a relatively small set of items across two or three different response formats (e.g., Carlisle, 2000, Goodwin et al., 2014; Nagy et al., 2003; Wilson-Fowler & Apel, 2015). Our analysis suggests that having a relatively large set of items tested across a diversity of task formats does not guarantee sensitivity to sublexical processing in morphological awareness tasks.

Implications for task construction and study design

Our goal in this study was to construct and validate an item level measurement of MA. Having done so, we have unlocked more questions than we actually answered. The implication for test design is that avoiding measurement error is related to having an item set that is broad enough within tasks to help capture differences in sensitivity to lexical and sublexical factors. The effects we observed in the current study cannot be seen in short and efficient tests that reify morphological knowledge into a latent construct, using CTT or IRT to create the smallest set of items possible to measure it on a set of subtests (Goodwin et al., 2017; Wilson-Fowler & Apel, 2015). These efficient tests are too narrow and are not linguistically informative. There is a need to build much broader and more nuanced tests that can be part of future explorations of how the properties of items interact with properties of individuals in crossed random and multilevel models. The morpheme counting task can help in this process because there is no need to create a carrier sentence for each item, simplifying the test construction process and shortening administration times.

Limitations and future directions

Our goal was to determine if morphological awareness could be measured at the word level in college students. Our findings are limited by the age and skill level of our participants. The morpheme counting task was designed to follow the protocol of syllable and phoneme counting tasks for beginning readers (Liberman et al., 1974). We are very interested in how the task operates developmentally, in younger readers with the same items; but this work was beyond the scope of the current experiments. Another point of interest is the contrast between the effects of phonological and orthographic shifts in the MCT task—we added these items for a relatively small sample of the MCT participants. Also with respect to the items, we did not explore differences between free and bound morphemes—English has many items that etymologically contain multiple morphemes that are not productively used in English (e.g., the monomorphemic English word *obituary* comes from the Latin *obitus* + the suffix *–arius/ary*). It is indisputable that better readers do encode this information in their lexicons and the role of this richer encoding in comprehension is interesting, but it was beyond the scope of initial task validation study. Finally, with an average RT of over 3 s, we did expect the weak relations between reaction time in MCT and other tasks, however we did not expect the lack of a significant correlation between RT and accuracy in MCT itself. Future work could examine how reaction time in MCT is related to variables that go beyond the decoding and word meaning measures that we employed, such as working memory.

Conclusions

In conclusion, the open set of derived words in English is massive. All studies that measure this complex knowledge with a handful of carefully chosen items risk the reification of an oversimplified view of morphology as a latent construct that can be mastered. The true target of measurement is the extent to which representations of morphology, phonology, orthography, and meaning are fully specified and interconnected (Perfetti, 2007; Perfetti & Hart, 2002). Creating the items within these tasks is challenging because of sentence, item, and person level constraints. MCT can simplify the process somewhat in that there are not additional demands of creating carrier sentences and additional variance from non-target items. The current study of item difficulty represents a first step in understanding how difficulty is related to vocabulary knowledge, word families, affix frequencies, and shifting.

Appendix 1

Items	Accuracy	Difficulty	*SEM*
Morpheme counting task items			
One morpheme items			
abandon	.68	− 0.86	0.23
absolute	.61	− 0.49	0.22
aluminium	.61	− 0.49	0.22
apparatus	.56	− 0.28	0.22
avocado	.80	− 1.61	0.26
barracuda	.61	− 0.54	0.22
obituary	.57	− 0.32	0.22
restaurant	.55	− 0.24	0.22
Two morpheme items			
considerable	.49	0.06	0.22
manageable	.77	− 1.43	0.25
marvelous	.63	− 0.63	0.22
stutterer	.60	− 0.45	0.22
suggestion	.70	− 1.00	0.23
traveler	.58	− 2.10	0.29
vigorous	.86	− 0.36	0.22
wanderer	.67	− 0.81	0.23
Three morpheme items			
dishonorable	.79	− 1.54	0.26
disinterested	.72	− 1.10	0.24
headers	.79	− 1.54	0.26
reconsideration	.47	0.14	0.22
renamed*	.75	− 1.26	0.24
tenths	.25	1.32	0.24

Items	Accuracy	Difficulty	*SEM*
traders*	.61	− 0.49	0.22
unreasonable	.89	− 2.37	0.32
Four morpheme items			
broadcasters	.68	− 0.86	0.23
encampments	.61	− 0.49	0.22
homecomings*	.46	0.18	0.22
landowners	.57	− 0.32	0.22
mockingbirds	.48	0.10	0.22
removals*	.11	2.35	0.31
statesmanship	.34	0.79	0.22
unbeknown	.13	2.17	0.30

Items with orthographic shifts are marked with *

Appendix 2

Additional shift items for MCT tas
Two morpheme shift items
abolition
alliance
courageous
criticize
designate
divisive
plurality
superiority
Three morpheme shift items
collisions
confidently
nationalist
simplifying
sprinklers
substantiated
vaccinated
worriedly

References

Anglin, J. M. (1993). Vocabulary development: A morphological analysis. *Monographs of the Society for Research in Child Development*, *58*(10), v-165. https://doi.org/10.2307/1166112.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., et al. (2007). The English lexicon project. *Behavior Research Methods, 39*(3), 445–459.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2018). *Fitting linear mixed-effects models using lme4*. arXiv:1406.5823 Stat [Internet]. 2014.

Berko, J. (1958). The child's acquisition of English morphology. *Word Journal of The International Linguistic Association, 14*(March), 150–177. https://doi.org/10.1080/00437956.1958.11659661.

Berninger, V. W., Abbott, R. D., Nagy, W., & Carlisle, J. (2010). Growth in phonological, orthographic, and morphological awareness in grades 1 to 6. *Journal of Psycholinguistic Research, 39*(2), 141–163.

Carlisle, J. F. (1988). Knowledge of derivational morphology and spelling ability in fourth, sixth, and eighth graders. *Applied Psycholinguistics, 9*(3), 247–266.

Carlisle, J. F. (2000). Awareness of the structure and meaning of morphologically complex words: Impact on reading. *Reading and Writing, 12,* 169–190. https://doi.org/10.1023/A:1008131926604.

Carlisle, J. F., & Fleming, J. (2003). Lexical processing of morphologically complex words in the elementary years. *Scientific Studies of Reading, 7*(3), 239–253.

Carlisle, J. F., & Kearns, D. M. (2017). Learning to read morphologically complex words. *Theories of Reading Development, 15,* 191–214. https://doi.org/10.1075/swll.15.11car.

Carlisle, J. F., & Stone, C. A. (2005). Exploring the role of morphemes in word reading. *Reading Research Quarterly, 40*(4), 428–449. https://doi.org/10.1598/RRQ.40.4.3.

Carson, R. J., & Beeson, C. M. L. (2013). Crossing language barriers: Using crossed random effects modelling in psycholinguistics research. *Tutorials in Quantitative Methods for Psychology, 9*(1), 25–41.

Chetail, F., Balota, D., Treiman, R., & Content, A. (2015). What can megastudies tell us about the orthographic structure of English words? *The Quarterly Journal of Experimental Psychology, 68*(8), 1519–1540. https://doi.org/10.1080/17470218.2014.963628.

Cho, S. J., Gilbert, J. K., & Goodwin, A. P. (2013). Explanatory multidimensional multilevel random item response model: An application to simultaneous investigation of word and person contributions to multidimensional lexical representations. *Psychometrika, 78*(4), f830–f855. https://doi.org/10.1007/s11336-013-9333-5.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*(4), 335–359.

Cooper, J. L. (2017). *Examining incidental vocabulary acquisition by person- and item-level factors in secondary students* (ProQuest Information & Learning; Vol. 77). Retrieved August 6, 2019 from https://ezproxy.mtsu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2016-53061-286&site=ehost-live&scope=site.

Cooper, J., Coggins, J., & Elleman, A. (2015). Morphology and mnemonic instruction: A comparison of two vocabulary methods on the vocabulary acquisition and reading comprehension of secondary students. In *22nd annual conference of the society for the scientific study of reading*. *Hapuna Beach, Hawaii*.

Derwing, B. L., & Baker, W. J. (1979). Recent research on the acquisition of English morphology. *Language Acquisition*, 209–223.

Ehri, L. C., & Robbins, C. (1992). Beginners need some decoding skill to read words by analogy. *Reading Research Quarterly, 27*(1), 12–26. https://doi.org/10.2307/747831.

Fowler, A. E., Liberman, I. Y., & Feldman, L. B. (1995). The role of phonology and orthography in morphological awareness. *Morphological Aspects of Language Processing*, 157–188.

Goodwin, A. P., Gilbert, J. K., Cho, S.-J., & Kearns, D. M. (2014). Probing lexical representations: Simultaneous modeling of word and reader contributions to multidimensional lexical representations. *Journal of Educational Psychology, 106*(2), 448.

Goodwin, A. P., Petscher, Y., Carlisle, J. F., & Mitchell, A. M. (2017). Exploring the dimensionality of morphological knowledge for adolescent readers. *Journal of Research in Reading, 40*(1), 91–117. https://doi.org/10.1111/1467-9817.12064.

Guo, Y., Roehrig, A. D., & Williams, R. S. (2011). The relation of morphological awareness and syntactic awareness to adults' reading comprehension: Is vocabulary knowledge a mediating variable. *Journal of Literacy Research, 43*(2), 159–183. https://doi.org/10.1177/1086296X11403086.

Habash, N., & Dorr, B. (2003). A categorical variation database for English. In *Proceedings of the North American association for computational linguistics* (pp. 96–102). Retrieved August 6, 2019 from https://clipdemos.umiacs.umd.edu/catvar/.

Haft, S. L., Myers, C. A., & Hoeft, F. (2016). Socio-emotional and cognitive resilience in children with reading disabilities. *Current Opinion in Behavioral Sciences, 10,* 133–141.

Kearns, D. M. (2015). How elementary-age children read polysyllabic polymorphemic words. *Journal of Educational Psychology, 107*(2), 364–390. https://doi.org/10.1037/a0037518.

Kieffer, M. J., Petscher, Y., Proctor, C. P., & Silverman, R. D. (2016). Is the whole greater than the sum of its parts? Modeling the contributions of language comprehension skills to reading comprehension in the upper elementary grades. *Scientific Studies of Reading, 20*(6), 436–454.

Kuo, L., & Anderson, R. C. (2006). Morphological awareness and learning to read: A cross-language perspective. *Educational Psychologist, 41*(3), 161–180.

Law, J. M., Wouters, J., & Ghesquière, P. (2015). Morphological awareness and its role in compensation in adults with dyslexia. *Dyslexia, 21*(3), 254–272. https://doi.org/10.1002/dys.1495.

Levesque, K. C., Kieffer, M. J., & Deacon, S. H. (2017). Morphological awareness and reading comprehension: Examining mediating factors. *Journal of Experimental Child Psychology, 160,* 1–20. https://doi.org/10.1016/j.jecp.2017.02.015.

Liberman, I. Y., Shankweiler, D., Fischer, F. W., & Carter, B. (1974). Explicit syllable and phoneme segmentation in the young child. *Journal of Experimental Child Psychology, 18*(2), 201–212. https://doi.org/10.1016/0022-0965(74)90101-5.

Linacre, J. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7,* 328.

Maag, L. K. (2007). *Measuring morphological awareness in adult readers: Implications for vocabulary development*. Gainesville: University of Florida.

Mahony, D. L. (1994). Sensitivity to word structure R&W.pdf. *Reading and Writing, 6*(1), 19–44.

McCutchen, D., & Logan, B. (2011). Inside incidental word learning: Children's strategic use of morphological information to infer word meanings. *Reading Research Quarterly*, *46*(4), 334–349.

Melnychuk, M., Banich, M., Milham, M., Atchley, R., Cohen, N., Webb, A., et al. (2013). NIH Public Access. *Scientific Studies of Reading*, *15*(1), 1–22. https://doi.org/10.1007/s11145-013-9426-7.

Nagy, W. E., & Anderson, R. C. (1984). How many {words} are {there} in printed {school} {English}? *Reading Research Quarterly, 19*(3), 304–330. https://doi.org/10.2307/747823.

Nagy, W., Berninger, V., Abbott, R., Vaughan, K., & Vermeulen, K. (2003). Relationship of morphology and other language skills to literacy skills in at-risk second-grade readers and at-risk fourth-grade writers. *Journal of Educational Psychology, 95*(4), 730–742. https://doi.org/10.1037/0022-0663.95.4.730.

Nunes, T., & Bryant, P. (2006). *Improving literacy by teaching morphemes*. London: Routledge.

Nunes, T., Bryant, P., & Bindman, M. (1997). Morphological spelling strategies: Developmental stages and processes. *Developmental Psychology, 33*(4), 637–649. https://doi.org/10.1037/0012-1649.33.4.637.

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading, 11*(4), 357–383. https://doi.org/10.1080/10888430701530730.

Perfetti, C. A., & Hart, L. (2002). The lexical quality hypothesis. *Precursors of Functional Literacy, 11,* 67–86.

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

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879.

Rubin, H. (1988). Morphological knowledge and early writing ability. *Language and Speech, 31*(4), 337–355.

Selby, S. (1972). The development of morphological rules in children. *British Journal of Educational Psychology, 42*(3), 293–299. https://doi.org/10.1111/j.2044-8279.1972.tb00722.x.

Spencer, M., Muse, A., Wagner, R. K., Foorman, B., Petscher, Y., Schatschneider, C., et al. (2015). Examining the underlying dimensions of morphological awareness and vocabulary knowledge. *Reading and Writing, 28*(7), 959–988. https://doi.org/10.1007/s11145-015-9557-0.

Stoet, G. (2010). PsyToolkit: A software package for programming psychological experiments using Linux. *Behavior Research Methods, 42*(4), 1096–1104.

Tighe, E. L., & Binder, K. S. (2015). An investigation of morphological awareness and processing in adults with low literacy. *Applied Psycholinguistics, 36*(2), 245–273. https://doi.org/10.1017/S0142716413000222.

Tighe, E. L., & Schatschneider, C. (2015). Exploring the dimensionality of morphological awareness and its relations to vocabulary knowledge in adult basic education students. *Reading Research Quarterly, 50*(3), 293–311. https://doi.org/10.1002/rrq.102.

Torgesen, J. K., Wagner, R., & Rashotte, C. (2012). *Test of Word Reading Efficiency: (TOWRE-2)*. Pearson Clinical Assessment.

Treiman, R., & Baron, J. (1981). Segmental analysis ability: Development and relation to reading ability. *Reading Research: Advances in Theory and Practice, 3,* 159–198.

Tyler, A., & Nagy, W. (1989). The acquisition of English derivational morphology. *Journal of Memory and Language, 28*(6), 649–667. https://doi.org/10.1016/0749-596X(89)90002-8.

Ulicheva, A., Harvey, H., Aronoff, M., & Rastle, K. (2018). Skilled readers' sensitivity to meaningful regularities in English writing. *Cognition*. https://doi.org/10.1016/j.cognition.2018.09.013.

Wilson-Fowler, E. B., & Apel, K. (2015). Influence of morphological awareness on college students' literacy skills: A path analytic approach. *Journal of Literacy Research, 47*(3), 405–432. https://doi.org/10.1177/1086296X15619730.

Woodcock, R. W. (2011). Woodcock reading mastery tests. American Guidance Service.

Yap, M. J., & Balota, D. A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory and Language, 60*(4), 502–529. https://doi.org/10.1016/j.jml.2009.02.001.