Published in final edited form as:

Dev Psychol. 2021 March ; 57(3): 361–375. doi:10.1037/dev0001153.

Automaticity as an Independent Trait in Predicting Reading Outcomes in Middle-School

Tanja C Roembke[*,1,2], **Eliot Hazeltine**[2], **Deborah K Reed**[3], **Bob McMurray**[2]

[1]Institute of Psychology RWTH Aachen University Jaegerstr. 17/19, D-52066 Aachen, Germany

[2]Dept. of Psychological and Brain Sciences University of Iowa, W311 Seashore Hall, Iowa City, IA 52242-1407, USA

[3]Iowa Reading Research Center, College of Education, University of Iowa, 103 S. Lindquist Center, Iowa City, IA 52242, USA

Abstract

Many middle-school students struggle with basic reading skills. One reason for this might be a lack of automaticity in word-level lexical processes. To investigate this, we used a novel backward masking paradigm, in which a written word is either covered with a mask or not. Participants (N = 444 (after exclusions); n_{female} = 264, n_{male} = 180) were average to struggling middle-school students from an urban area in Eastern Iowa that were all native speakers of English and were roughly equally from grades 6, 7 and 8 (average age: 13 yrs). Two-hundred-fifty-five students qualified for free or reduced-price lunch, a proxy for economic disadvantage. Participants completed different masked and unmasked task versions where they read a word and selected a response (e.g., a pictured referent). This was related to standardized measures of decoding, fluency and reading comprehension. Decoding was uniquely predicted by knowledge (unmasked performance), whereas fluency was uniquely predicted by automaticity (masked performance). Automaticity was stable across two testing points. Thus, automaticity should be considered an individually reliable marker/reading trait that uniquely predicts some skills in average to struggling middle-school students.

Keywords

Automatic word recognition; reading; fluency; backward masking

Reading is fundamental for academic success (National Institute of Child Health and Human Development, 2000). However, many students' reading education is insufficient, and they struggle with basic word-level reading skills like word recognition, decoding and fluency. These deficits are persistent. Even in middle-school, roughly half of struggling readers show word-level deficiencies (Cirino et al., 2013; Nippold, 2017).

Much of the variance in reading comprehension is explained by a combination of word-level orthographic skills and higher-level oral language skills like listening comprehension (e.g.,

[*]corresponding author tanja.roembke@psych.rwth-aachen.de, Phone: +49 (0) 241-80-93993.

Catts, Fey, Zhang, & Tomblin, 1999; Cutting & Scarborough, 2006). This is consistent with the Simple View of Reading (Gough & Tunmer, 1986), in which efficient word reading allows children to use written input to access higher-level oral language skills for comprehension. As a result, given the large number of older readers that struggle with word-level lexical skills, it is crucial to understand word recognition in low-performing readers beyond elementary school. The present study builds on recent work highlighting the importance of automatic word-level processes ("automaticity") in this population. We use this term to mean not only lexical processes like word recognition (mapping print to meaning), but also other processes, like mapping between print and sound, that support word recognition. We contrast *automaticity* with students' *knowledge* of the reading system (the letter/sound mapping, sight words).

Automaticity and Reading

If children functionally cannot read words rapidly, they may not have the resources to deploy higher-level language skills. For instance, Cutting and Scarborough (2006) found that word reading speed accounted for unique variance in reading comprehension after controlling for decoding and oral language. Moreover, some intervention studies find that students can improve decoding without achieving adequate fluency (e.g., de Jong & van der Leij, 1999; Torgesen & Hudson, 2006). This suggests that knowledge alone is insufficient for fluent reading.

Word reading is a product of the direct mapping of print to meaning (word recognition) but also the mapping of print to sound (Seidenberg, 2005). Word reading speed is often conceptualized in terms of *automaticity*. This may emerge in multiple ways in this complex system. At the broadest level, automatic processes are executed quickly, effortlessly, and without awareness (Logan, 1997). Several researchers propose that automatic word-level processes are important for outcomes like fluency and comprehension (Klauda & Guthrie, 2008; LaBerge & Samuels, 1974; Perfetti, 1985; Rasinski et al., 2005). By these accounts, automaticity allows for fluent reading, freeing resources for comprehension (Walczyk, 2000). Conversely, non-automatic word reading creates a "bottleneck"—even students who have knowledge of the letter-sound mappings and know the orthographic pattern of many words may not be good readers because they cannot access words quickly enough. This may be particularly the case for older children who possess the relevant knowledge but continue to struggle.

There are three gaps in our understanding of how automaticity relates to word reading. First, there has been little consideration of the contribution of automaticity in different parts of the reading system. Second, even though automaticity is considered necessary for fluency and comprehension, the developmental relationship between these skills remains unclear. Finally, there is only limited empirical evidence for automaticity (distinct from general processing speed) as a unique contributor to reading outcomes.

The best evidence for a unique role of automaticity comes from Lovett (1987). She tested three groups of children (8–13 year-olds): *accuracy-disabled*, *rate-disabled*, and *fluent* children. Accuracy-disabled children struggled with both decoding and fluency, whereas rate-disabled children could decode words accurately but scored low in reading speed. Rate-disabled children were slower than the accuracy-matched control group at single word recognition. Moreover, rate-disabled children showed decreased decoding accuracy and comprehension when words were presented in context, suggesting that fluent reading is necessary for sentence comprehension.

These results support the idea that at the level of word reading, competent readers require at least two partially independent skills: knowledge and automaticity. However, these results are based on a small set of children with clinically identified deficits. Thus, they may not generalize to a wider range of children who struggle with automaticity in less severe ways. Most importantly, these groups were selected to be different in key characteristics (e.g., fluency, decoding accuracy). As a result, it is unclear how these reading skills relate to each other in a more representative sample including children with average reading abilities.

Potential Loci of Automaticity

Studies like Lovett (1987) and Cutting and Scarborough (2006) focus on automaticity at a functional level—how fast can children read words, and how many can they read in a minute (fluency). However, speed may arise by several mechanisms. Theoretical models suggest children can access meaning from print with two pathways. A word's meaning can be activated by activating its phonology first (orthography [O] → phonology [P] → semantics [S]) as in true decoding (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Harm & Seidenberg, 2004). But a known word's meaning can also be activated directly from its written form (orthography→semantics). Automaticity may thus be achieved in multiple ways. As the orthography → semantics path is direct, switching to this route may increase speed (Harm & Seidenberg, 2004). In dual-route models, automaticity can only be achieved for known words (Coltheart et al., 2001; Share, 1995, 2008; Tamura, Castles, & Nation, 2017), as decoding (the O→P pathway) is a serial process. Automaticity emerges as the child switches from one path to another, but processing within a path is relatively constant.

In contrast, in connectionist models of reading (e.g., Seidenberg & McClelland, 1989), automaticity can exist in both direct and indirect paths, and these pathways could develop independently. For example, one person could activate sight words automatically from orthography but be unable to map text to print automatically (e.g., due to a phonological deficit or lack of decoding instruction). Another person may have less well-organized O→S mappings. Thus, automaticity may develop within either or both pathways. A critical test is whether automaticity is possible in the O→P pathway. This can be assessed with nonwords where the semantic pathway is not available.

A New Way to Measure Automaticity: Backward Masking

Isolating the speed of word recognition from other cognitive processes is difficult for several reasons. First, differences in reading speed could derive from differences in knowledge—a child who does not know all the letter/sound mappings will not read quickly. This is why Rapid Automatized Naming (RAN), which stresses well-known words (or even just letters and numbers), is useful for predicting reading outcomes (e.g., Denckla & Rudel, 1974; Lervåg & Hulme, 2009). However, as RAN tasks are limited to highly familiar words, they cannot measure automaticity across the full range of letters, sounds, and words. Second, people may show different reaction times in word recognition tasks because of differences in cognitive factors that are independent of reading. They may vary in their decision speed or how rapidly decisions are mapped onto action. Measures like fluency or RAN are also susceptible to differences in speech rate. Third, it can be challenging to divorce the automaticity of specific skills from general changes in processing speed over development (Kail & Hall, 1994). Finally, cognitive psychology has long recognized a trade-off between speed and accuracy—people can respond quickly (at the expense of accuracy) or more deliberately (and accurately). Thus, some of the variance in response time may reflect students' setting of the speed/accuracy trade-off. This could affect different components of recognition, response selection, and motor planning (Dickman & Meyer, 1988). Thus, despite theoretical arguments that automaticity should capture unique variance in word recognition, there is little evidence because of challenges measuring word reading speed.

Backward masking can overcome many of these issues. In this paradigm, words are presented for a short period of time (e.g., 80 msec) and covered by a visual mask (e.g., ####). To respond accurately, participants must rapidly map the visual input onto a more durable representation (e.g., phonological or semantic codes)—once the mask covers the word, it is no longer available in the input or in visual memory. That means a correct response is only possible if activation spreads automatically from the orthographic representation to the phonological or semantic codes before the appearance of the mask.

Backward masking has a history in studies of visual word recognition where it has been used to isolate the earliest stages of processing where priming or inhibition can occur without awareness (e.g., Forster & Davis, 1984; Reicher, 1969; Wheeler, 1970). It is analogous to "flash" testing, where experimenters and educators present cards with written words for about half a second before removing them to test children's automaticity (Morris et al., 2013, 2012).

Backward masking can divorce automaticity from knowledge. Performance with unmasked text can measure knowledge (i.e., does the child recognize the word or letter string), and assess general task demands (e.g., attention), whereas performance with masked words requires all these skills but also requires automaticity. By assessing both simultaneously, the unique contribution of automaticity can be isolated. Masking also isolates word recognition speed from general speed of processing and the speed/accuracy trade-off. The child has unlimited time to respond; thus, the real measure is whether the reading system can extract enough information in a short period. Unlike RAN, backward masking can be used with any stimuli (including nonwords), to isolate automaticity on the orthography → phonology paths. Thus, masking overcomes limitations of prior measures and when combined with unmasked versions of the same tasks, is well-suited to assess the automaticity of word-level processes.

Following this line of reasoning, Roembke et al. (2019) used masking to ask whether automaticity (performance in tasks with masked stimuli) predicted reading outcomes *over* *and above* students' knowledge (unmasked performance on the same tasks). This was intended to isolate automaticity as a uniquely predictive trait. We studied middle-school students with below-average to average reading skills' fluency, a population that often has the relevant knowledge of letters, GPC regularities, and words but can struggle to read fluently. Nevertheless, this population has received relatively little attention in reading research (Cirino et al., 2013).

Masking was implemented in three tasks designed to maximize reliance on different reading pathways (Table 1). In *Find the Picture*, participants saw a written word along with four images and selected the corresponding picture (emphasizing automatic activation of semantics directly from orthography, the O→S pathway). In *Find the Rhyme*, students matched a written word to one of eight potential rhyming words (emphasizing automatic activation of phonology, the O→P pathway). Finally, in *Verify*, an auditory word was presented and the participant indicated whether it matched a written word (likely tapping both pathways). Decoding, fluency and comprehension were assessed as outcome measures.

Performance in the masked tasks predicted fluency over and above the same unmasked task. This was particularly strong for *Find the Picture*, emphasizing direct semantic mappings, and for nonwords in *Find the Rhyme*. This latter finding suggests automaticity may characterize the orthography→phonology path as well. For decoding, masked performance accounted for little unique variance, but unmasked performance had a unique effect. These results suggest automaticity may be a unique predictor of specific reading skills (fluency), and raise the possibility that many struggling readers may be limited by a lack of automaticity (Lovett, 1987).

Roembke et al. (2019) leaves a number of questions unanswered. First, there were insufficient subjects for multivariate analysis combining all tasks—each task was examined individually. A multivariate analysis may more definitively identify the variance in outcomes that is uniquely related to automaticity in each task. This can provide insight into the contributions of knowledge or automaticity in each reading pathway and inform interventions to improve word recognition.

Second, the study used only monosyllabic words. It is not clear if predictive relationships between reading outcomes and automaticity extend to multisyllabic ones, which are lower in frequency and acquired later (Ferguson & Farwell, 1975). Although previous research shows that the recognition and naming of multisyllabic and monosyllabic words shows similar patterns (Jared & Seidenberg, 1990), multisyllabic words also require additional processes like syllabification, and invoke different GPC regularities (Bruck, 1990). Supporting this, adult readers with dyslexia have more difficulty reading multisyllabic words than age-matched control subjects, and instead perform similarly to reading-matched controls (Bruck, 1990). Similarly, the effect of word-length varies as a function of reading skill: Better readers are less likely to be affected by length (Manis, 1985). However, the degree to which reduced proficiency with multisyllabic words is due to an inability to activate them automatically is unclear. Nevertheless, it is possible that multisyllabic words may be more diagnostic of automaticity deficits, as a small slow-down in reading a few letters may cascade across multiple letters.

Finally, even though Roembke et al. (2019) showed that automaticity uniquely predicted fluency, it was unclear whether masking offered an individually reliable estimate. That is, if we tested participants again, would the effect of masking be similar for each participant? If yes, this would be evidence for automaticity as a unique dimension underlying readers' performance. This could in turn suggest automaticity as a useful target for assessments or potentially intervention.

The Present Study

The current study sought to examine predictive relationships between automaticity in word-level lexical processes and reading outcomes. This was a significant extension of Roembke et al. (2019), with an almost ten-fold increase in sample size, a much larger range of items, and a test/retest reliability assessment. This study was embedded in a validation and development study of a larger new assessment of middle-school reading (Iowa Assessment of Skills and Knowledge [iASK], Foundations in Learning, Inc., 2009). Our study focused on struggling to average readers. This was consistent with the goal of the broader assessment development and validation, but also with our scientific goals, as factors that predict above average reading were deemed less important to improving educational outcomes (Sorensen, 2019). Thus, students were identified as average to below-average readers, based on a state-wide measure of reading comprehension (in the 10th to 60th percentile), resulting in a large though restricted range that included both typical and struggling readers.

The validation study consisted of two independent waves of over 200 students. Roughly half the students in Wave 2 completed the test again a month later to estimate reliability. As in Roembke et al. (2019), we used three tasks (*Find the Picture*, *Find the Rhyme*, *Verify*), each designed to differentially emphasize the two major reading pathways. Each task was used in both masked and unmasked form and related to a variety of outcomes (comprehension, fluency and decoding). Stimuli consisted of words and nonwords, both monosyllabic and multisyllabic.

We addressed the following questions:

1. What are the respective contributions of automaticity and knowledge to middle-schoolers' decoding, fluency, and reading comprehension? We hypothesize a predictive relationship between automaticity and fluency, but not decoding when controlling for knowledge. Consistent with Roembke et al. (2019) and Lovett (1987), we hypothesized knowledge would uniquely predict decoding, not fluency, when controlling for automaticity.
2. Does automaticity in nonword processing account for unique variance in outcomes? This would constitute strong evidence that orthographic-phonemic mappings can be automatic.
3. Does the pattern of linkage between automatic and non-automatic processing and standardized reading outcomes hold for multisyllabic words?
4. What are the unique contributions of automaticity in each task to outcomes?
5. Is automaticity a stable trait that remains reliable across different time points?

Methods

Participants

Participants consisted of middle-school students (Grades 6–8) whose scores on the previous year's state summative reading test (a silent reading comprehension measure) was between the 10th and 60th percentile. As part of a University of Iowa Institutional Review Board approved protocol (IRB# 201507769, *iASK*), informed consent was given by parents, and students completed a group administered assent protocol. Students were recruited in two waves for two independent tests on consecutive years.

Wave 1 participants were 240 students (n_{female}= 139, n_{male} = 101; average age: 13;0, SD = 11 months). Students were recruited from four middle-schools in Cedar Rapids, an urban area in Eastern Iowa. Roughly equal numbers came from each grade (6th: n = 85; 7th: n = 83; 8th: n = 72). Nine were excluded, as they completed less than 50% of the trials; one was excluded because s/he did not complete all standardized assessments. This left 230 in the final dataset. Participants were all native English speakers. Twenty-one students qualified for special education, and eight qualified for a 504 plan. None of the students had an intellectual disability, but other disability categories were unknown due to the state's non-categorical policy for special education. One-hundred-eighteen of the remaining students qualified for free or reduced-price lunch, a proxy for economic disadvantage.

Wave 2 participants consisted of 241 students (n_{female} = 145, n_{male} = 96; average age: 12;8 years, SD = 15 months; 6th: n = 90; 7th: n = 77; 8th: n = 74) from four middle-schools in the same district (three of which were also part of Wave 1). None of the Wave 2 students were tested in Wave 1. Eight students were not included because less than 50% of their data were available; 4 were excluded because of non-compliant behavior; 14 due to a lack of English proficiency which was not detected during screening; one student dropped out. This left 214 participants. Thirty-nine students qualified for special education, and nine qualified for a 504 plan. None of the students had an intellectual disability. One-hundred-thirty-seven qualified for free or reduced-price lunch.

A subset of Wave 2 participants (n = 127)[1] completed the same procedures one month later (though with partially distinct sets of items, and with a new random assignment of items to tasks). No standardized tests were administered at this time point. Participants were pseudorandomly selected to be included in the reliability assessment, so that the resulting reliability sample had a similar distribution to the entire Wave 2 sample in gender and state reading scores (Table 2). Out of the reliability sample, five children were excluded because they participated in less than half of all trials, resulting in 122 participants in reliability analyses (n_{female} = 69, n_{male} = 53; average age: 13;11 years; 6th: n = 44; 7th: n = 41; 8th: n = 37).

[1] This number does not include any student that was excluded from the main analysis of Wave 2 at time point 1.

Standardized Tests

Two subtests of the Woodcock Johnson Reading Mastery Test, 3rd edition assessed **word-level decoding** (Woodcock, McGrew, & Mather, 2001). In these tests, students read aloud familiar words (Word Identification, WRMT_ID), or low frequency words and nonwords (Word Attack, WRMT_AT). Items increased in difficulty. Both were untimed and administered individually. **Fluency** was assessed with the Texas Middle School Fluency Assessment (TMSFA; Francis et al., 2010). Students read as much of a passage as they could in one minute and were instructed to do their “best reading.” Each student completed three separate passages. Performance was scored based on numbers of words read correctly per minute and converted to an equated score to account for passage difficulty. **Reading comprehension** was assessed with the Gates-MacGinitie Reading Test, 4th edition (GMRT; MacGinitie, MacGinitie, Maria, & Dreyer, 2000), group administered. Students had 35 min to read a series of short passages and respond to multiple-choice questions. Finally, **oral vocabulary** was assessed with the Peabody Picture Vocabulary Test, 4th edition (PPVT; Dunn & Dunn, 2007). Students heard a word and identified the matching picture from an array of four pictures. The PPVT is untimed. It was used as a covariate, not an outcome measure, to account for general oral language abilities.

All tests were conducted on both waves[2]. Standard scores were calculated for the WRMT and PPVT. For the GMRT, standard scores are not available (only percentiles), so percentiles were converted to standard scores. For the TMSFA, no standard scores were available; scores were converted to an equated score and averaged across the three passages. Students completed the group-administered tests first in groups of about 20–30 students. For individual testing, students were assigned to one of four trained testers. Testers spread out to the corners of the room to minimize distraction; each tester started with a different test, so students in the same room did not complete the same test at the same time.

Overall Design

Experimental tasks were delivered via a custom program over the Internet as part of a developmental version of iASK (Foundations in Learning, Inc., 2009). In Wave 1, participants completed 1,200 trials over four days (the last day was for make-up testing). Six hundred trials were predefined as relevant to the research questions here. In Wave 2, participants completed 708 trials with 456 intended for these questions. The rest of the trials were intended for separate research questions and not analyzed. Participants were not aware which trials contributed to which study. Students did not always complete all trials, and only data of participants who completed more than 50% of relevant trials were used.

Between Waves 1 and 2, we modified the tasks and items based on focus-group feedback from students and teachers about the tasks (Reed, Martin, Hazeltine, & McMurray, 2019), and an Item Response Theory (IRT) analysis to exclude items whose estimated difficulty was easier or harder than expected (leading to a reduction of trials in Wave 2, while maintaining equally strong measures). The analyses reported here were not conducted until

[2] Wave 1 students were also tested on several other tests of reading vocabulary, syllabification, silent fluency, working memory and sentence comprehension. These were collected for different questions and were not analyzed as part of the present project.

both waves of data collection were complete. In this IRT analysis, a series of logistic mixed effects models were used to estimate item level difficulty (intercept) while accounting for subject level ability (an intercept of subject) and various properties of the items (e.g., spelling complexity, word versus nonword) and task. From these analyses we chose a subset of items whose estimated intercepts were near 0 (i.e., items that were about as difficult as expected given the fixed factors and the subjects' abilities). This did not include any analysis of the standardized outcome measures. Thus, item selection was blind to our hypotheses.

Six base lists controlled which words were assigned to which task. In Wave 1, there were 1,049 unique items and, in Wave 2, there were 1,248 unique items. In each list, about 40% of items were shared with another list (though items appeared in different tasks across lists); all items were used in at least two lists. Four versions of each list were created. These were constructed by crossing two factors: first, we constructed a matched, counterbalanced version in which the items that appeared in the masked condition for one list were presented in the unmasked version and vice versa. Second, we crossed the masking factor with the *Verify* task (see below). Thus, subjects were randomly assigned to one of 24 lists.

The experiment consisted of three base tasks (*Find the Picture*, *Find the Rhyme*, *Verify*)[3].The base form of each task included monosyllabic words. *Find the Picture* and *Verify* each had multisyllabic versions in which items were either 1, 2, or 3-syllable words (interleaved). There was not a multisyllabic version of *Find the Rhyme* due to item limitations. Each of these five tasks appeared in both masked and unmasked forms. In both the base and multisyllabic versions of *Find the Rhyme* and *Verify*, items consisted of both words and phonetically regular nonwords (randomly interleaved). Nonwords were not possible for *Find the Picture.*

At the beginning of a session, students were given a choice of task versions from a colorful selection screen. They chose a task version and then completed a block of 16–20 trials. Each screen contained an assortment of possible task versions (both tasks for this study, and tasks for the larger validation study) that did not repeat. When the student finished all the task/versions on a screen, they saw a new screen with a new set of choices. Thus, they completed multiple blocks of each task version, interleaved across testing.

For Wave 1, blocks were 20 trials and participants completed three iterations of each task version (3 repetitions × 20 trials × 5 tasks [*Find the Picture/Multi*, *Verify/Multi, Find the Rhyme*] × 2 masked/unmasked). For Wave 2, all tasks but one included 16 trials per iteration with three repetitions (3 repetitions × 16 trials × 4 tasks × 2 masking conditions). For the monosyllabic version of *Find the Picture*, there were 18 trials per block and two repetitions (2 repetitions × 18 trials × 2 masking conditions). *Find the Picture* needed fewer trials, as no nonword trials were possible. Within the trials for a specific task × masking cell, there were an equal number of words and nonwords, and an equal number of items by the class of vowel and consonant (e.g., digraphs, long vowels, etc.), and by number of syllables (1, 2, or 3).

[3] Correlations among all task versions are included in the online supplement S3.

Experimental Tasks

Each trial began with a "start button" (a blue triangle inside a square). The start button ensured participants were attending to the right location when the stimulus was presented. After clicking the button, students saw the target. In the unmasked version, the target appeared until a response was made. In the masked version, the target was covered with the mask after 80 msec. The mask was present until participants responded. Participants received no feedback. The start button for the next trial was presented 1,600 msec after the response. Trial order within a task was random, and the arrangement of response options on the screen was randomly determined.

Masking was consistent within a block of trials, and the assignment of items to a masked or unmasked version was counterbalanced across participants. For *Find the Rhyme* and *Verify*, real words and nonwords were interleaved within a block (Table 1). If a target was a word, all response options were words; for nonword targets, all responses were nonwords. Multisyllabic versions of *Find the Picture* and *Verify* were conducted in separate blocks from the monosyllabic versions. Each block contained roughly equal numbers of 1-, 2-, and 3-syllable words.

Find the Picture.—In this task, four response pictures (the target and three foils) were arranged horizontally. Participants clicked on the image that matched the target. Foils were close matches of the target in orthography and phonology (e.g., COMB as a foil for COAST). Semantic competitors (e.g., BEACH for COAST) were not used to avoid confusion.

Find the Rhyme.—In *Find the Rhyme*, students selected the rhyme of the target from eight written response options. Approximately 60% of the correct rhymes used the same orthographic vowel as the target word (e.g., target: MAIL; rhyme: SAIL). The remaining rhymes used a different letter string (e.g., target: MAIL; rhyme: SALE). These forced participants to attend to phonological similarity, not orthographic overlap. Two of the seven foils began with the same consonant(s) and vowel pronunciation as the target word (e.g., target: MAIL; competitors: MAZE and MAIN); two used the same beginning and ending consonant(s), (e.g., target: MAIL; competitors: MULE and MEAL); two overlapped with the target's vowel pronunciation only and also often matched a consonant with the rhyme (e.g., target: MAIL; competitors: SAME, SANE); the last foil could be matched on any other dimension (e.g., target: MAIL; competitor: RAIN). In a few cases, one or more of these categories of foils was not possible and a competitor that matched one of the other categories was added.

Verify.—In *Verify*, participants heard a spoken item and saw a written target. On 50% of the trials they matched. The task was to indicate whether the spoken and written stimulus matched by clicking a red button (mismatch) or a green one (match). On mismatch trials, the auditory stimulus was either a phonological or orthographic competitor. It could either overlap with the target item at onset (e.g., target: COAST; competitor: COACH), offset (e.g., target: RAIL; competitor: FAIL), or another location (e.g., target: FREAK; competitor: FLAKE).

Reliability of tasks.—Cronbach's alpha for each task was calculated using the Wave 1 data. The unmasked version of *Find the Picture*'s average alpha was 0.68, whereas its masked version showed higher reliability (α = 0.80). *Find the Rhyme* (unmasked: α = 0.94; masked: α = 0.93) and *Verify* (unmasked: α = 0.73; masked: α = 0.80) showed high reliability. For multisyllabic task versions, the unmasked version of *Find the Picture* had an alpha of 0.68, and its masked version had an alpha of 0.85. Reliability was also high for multisyllabic *Verify* (unmasked: α = 0.80; masked: α = 0.75). Task reliability was likely lower for *Find the Picture* (particularly the unmasked version) due to high overall performance, resulting in little variability across items.

Items and Stimuli

Items.—Items were selected to be known by middle-school students and were evaluated by a group of six (the authors and two curriculum developers at *Foundations in Learning*). They were selected to be appropriate for middle-school students (e.g., non-violent) and to be reflective of written text they might read. Half of the items included at least one consonant cluster or digraph. We selected an equal number of items within 7 vowel classes: short vowels (e.g., A as in CAT), long vowels (e.g., vowels with a silent E such as I_E as in MICE), secondary or irregular pronunciations of short and long vowels (e.g., O_E as in LOVE or I in KIND), vowel digraphs (e.g., EE as in BREED), secondary or exception digraphs (e.g., EA as in BREAD), and diphthong and R-controlled vowels (e.g., OI as in BOIL, AR as in BARK). Thus, approximately 4/5 of items included regular vowels. Nonwords were phonologically regular pseudowords (e.g., LOIF), constructed by selecting similar sound sequences as real words. Nonwords never included morpheme-like substrings or irregular (secondary or exception) vowels. When assigning items to tasks and blocks, items were balanced by consonant type and vowel class. Vowel and consonant classes were separately balanced within words and nonwords.

Auditory stimuli.—Auditory stimuli were clearly spoken by a female speaker at a slow rate. They were recorded using a Kay Elemetrics 4300B A-to-D system at 44,100 Hz in a sound attenuated room. Multiple exemplars of each word were recorded and the clearest was selected. These were then edited to eliminate any clicks or distracting elements and to include 50 msec of silence at the beginning and end of the recording.

Visual stimuli.—The images used in *Find the Picture* were colored line drawings. They were developed using a standard lab protocol (Roembke et al., 2019) in which multiple candidates for a given word were downloaded from a commercial clip art database. These were then viewed by a small focus group to identify the best depiction of that word. Finally, they were edited to ensure prototypical color and orientation, remove distracting backgrounds, and so forth.

Procedure

For both waves, daily testing lasted approximately 45 min each, and students were supervised throughout testing by an experimenter who assisted with computer problems, answered questions, and guaranteed low distraction levels. Time on experimental tasks differed across days, with time allocated to set-up and deliver verbal instructions. Students participated for approximately four days in Wave 1 and three days in Wave 2.

Results

Table 2 shows an overview of the standardized assessments. In both waves, scores were below normative average. This was expected given recruitment criteria.

We first asked whether masking decreased accuracy. Next, we used a communality analysis to identify the unique contributions of masked and unmasked performance. Third, we examined nonword trials to isolate the orthography-to-phonology pathway. Finally, we examined test/retest reliability. Analyses were done separately for Waves 1 and 2, with identical models.

The Effect of Masking

To document that masking had an effect on accuracy, we used a series of repeated-measures ANOVAs. These were conducted separately for each task version, as these had separate chance levels (*Find the Picture*: 4 response options; *Find the Rhyme*: 8; *Verify*: 2), and we were not interested in comparing performance across tasks. For *Find the Rhyme* and *Verify*, two within-subject factors were included: masking (masked or unmasked) and item-type (word or nonword). For *Find the Picture*, only masking was examined (there were no nonwords). Accuracy was transformed with the empirical logit transformation.

Figure 1 and Table 3 summarize the results. For each task, there was a significant effect of masking with masked performance significantly lower than unmasked. The main effect of item-type was always significant, as accuracy was lower for nonwords than words. The masking × item-type interaction was not significant except in *Find the Rhyme (monosyllabic)* in Wave 1, where the effect of masking was more pronounced for nonwords. These analyses indicate that masking had the intended effect of increasing pressure on the reading system. They also show that nonwords were harder. This was expected, as participants cannot use previous knowledge or sight word processes to identify the item. Finally, these analyses reveal a similar pattern of processing for tasks using mono- and multisyllabic words.

Unique contributions of Automaticity and Knowledge: Statistical Framework

We used a series of communality analyses to investigate the contribution of automatic reading over and above knowledge of the words and letter/sound mappings. This allowed us to estimate the shared and unique variance in each outcome associated with the unmasked and masked task versions. Analyses were run separately with each reading outcome measure as a predictor. The analyses assume that performance on the unmasked task version captures relevant knowledge, whereas the masked version captures the contribution of automaticity. We also estimate shared variance (e.g., general reading ability, ability to complete the tasks).

Communality analyses used two hierarchical regressions. First, we isolated the contribution of masked performance. On the first step, PPVT was entered to control for oral vocabulary (Jenkins, Fuchs, van den Broek, Espin, & Deno, 2003; Oslund, Clemens, Simmons, Smith,

& Simmons, 2016) to a outcome. In the second step, unmasked task versions were added, and ΔR^2 was recorded. All three relevant tasks (e.g., *Find the Picture*, *Find the Rhyme*, and *Verify*) were entered simultaneously to capture their joint contributions. Finally, the masked versions of all three tasks was entered. Here, ΔR^2 represents the variance in the outcome uniquely associated with automaticity. We next repeated this regression, but this time masked performance was entered on the second step and unmasked on the third. This identified the unique variance associated with knowledge. Shared variance was computed by subtracting unique contributions of masked and unmasked from their total contributions.

Analyses were conducted separately for monosyllabic tasks (Question 1), nonword trials (Question 2) and multisyllabic tasks (Question 3). Each wave was analyzed separately, but presented together, separated by outcome measure. Regression analyses were implemented in R (version Ri386 3.5.3; R Core Team, 2013).

Analysis 1: Monosyllabic Items

Figure 2A and Table 4 show the results of the communality analysis for monosyllabic words. Overall, regressions showed strong effects with the experimental tasks accounting for between 40% and 50% of the variance in fluency, about 38% of the variance in decoding, and 20% to 30% of the variance in comprehension. In both waves, experimental tasks were most predictive of fluency. This was surprising because all the tasks targeted isolated word-level processes, and none involved oral reading. We also saw a large increase in effect size between Waves 1 and 2 for fluency and comprehension, with increases of more than 10% of the variance. This probably reflects the improved items (as other changes were cosmetic).

Across the board, most of the outcome variance was shared between masked and unmasked tasks (light gray portion of the bars in Figure 2A). However, there were large unique effects of both masked and unmasked performance for some outcomes. Masked performance tended to be more important for fluency, whereas unmasked performance better predicted decoding skills. Oral vocabulary accounted for less than 2% of the variance in fluency and decoding and about 6% of variance in comprehension (see Figure 2A).

Both the unmasked and masked task versions accounted for significant unique variance in **fluency** (TMSFA). However, during both waves, masked task versions accounted for roughly twice as much variance in fluency in Wave 1 (unmasked $\Delta R^2 = 0.030$; masked $\Delta R^2 = 0.065$) and Wave 2 (unmasked $\Delta R^2 = 0.037$; masked $\Delta R^2 = 0.067$). These findings indicate that fluency reflects a larger component of automaticity than knowledge. In both Waves 1 and 2, unmasked tasks predicted unique variance in **decoding** (WRMT; Wave 1: $\Delta R^2 = 0.112$; Wave 2: $\Delta R^2 = 0.060$), whereas masked versions did not (Wave 1: $\Delta R^2 = 0.008$; Wave 2: $\Delta R^2 = 0.015$). This pattern suggests that performance on unmasked tasks was more critical than masked ones in predicting decoding and that automaticity may not be necessary for better decoding, at least as measured by unspeeded tests like the WRMT.

There was no consistent pattern for **comprehension** (GMRT) across waves. In Wave 1, masked task versions contributed unique variance ($\Delta R^2 = 0.059$), whereas unmasked did not. In contrast, in Wave 2, the unmasked task versions predicted unique variance ($\Delta R^2 = 0.030$)

but the masked ones did not. This inconsistency may derive from the fact that our tasks simply accounted for less overall variance in comprehension than other outcomes.

Analysis 2: Automaticity in Nonwords

We next investigated the relationship between automaticity and reading outcomes when experimental predictors derived from only nonword trials. As nonwords can only be evaluated by directly mapping orthography to phonology (decoding), automaticity in nonword processing would be further evidence that automaticity is not confined to direct O→S mappings or to words that have been learned as a whole (Share, 1995, 2008). The same communality strategy was used, but only nonword trials of *Find the Rhyme* and *Verify* contributed.

Figure 2B and Table 4 show results. Analyses accounted for nearly as much overall variance in outcomes as the prior analysis. Decoding (~38% of the variance) showed no reduction, whereas fluency (30–40% of the variance) and comprehension (13–28%) were reduced by roughly 10% of the total variance. As before, most variance was shared between masked and unmasked versions, but again there were moderate unique effects of each for some tasks.

In Wave 1 performance in the masked task versions accounted for significant variation in **fluency**, whereas unmasked versions did not (masked $\Delta R^2 = 0.093$; unmasked $\Delta R^2 = 0.016$). For Wave 2, both unmasked and masked tasks accounted for unique variance in fluency, but unique variance was slightly higher for the unmasked task versions (unmasked $\Delta R^2 = 0.064$; masked $\Delta R^2 = 0.049$). This finding was surprising given the results of the analysis of all trials, where masked versions were consistently more predictive of fluency. Unmasked task versions were associated with significant unique variance in **decoding** in Wave 1 ($\Delta R^2 = 0.119$) and Wave 2 ($\Delta R^2 = 0.069$). In addition, in Wave 2, masked task versions also predicted variance in decoding (ΔR2 = 0.030); this was not true for Wave 1 (ΔR2 = 0.014). These results are consistent with the analysis of all trials, suggesting that decoding is better predicted by knowledge than automaticity. Unmasked performance contributed unique variance in **comprehension** for Wave 2 ($\Delta R^2 = 0.054$) but not Wave 1 ($\Delta R^2 = 0.017$). In addition, masked performance was not significant in either wave and had small effects (ΔR^2 both waves ≤ 0.020).

In summary, unique variance in outcomes was associated with masked performance in nonwords. This was particularly true for fluency and, less so, for decoding. These results are consistent with the hypothesis that automaticity is not limited to the direct O→S pathway but can also exist in O→P mappings as assessed by nonword performance.

Analysis 3: Automaticity in Multisyllabic Words

We next conducted the same communality analyses on the two multisyllabic tasks (*Find the Picture* and *Verify*). The results of these analyses are summarized in Figure 2C and Table 4. Overall these analyses showed highly similar effect sizes as for monosyllabic words for fluency (40–50% of the variance) and comprehension (18–30% of the variance) and reduced (though still large) effect sizes for decoding (~28% of the variance).

In both waves, the masked task versions (Wave 1: $\Delta R^2 = 0.100$; $\Delta R^2 = 0.220$) accounted for more unique variance in **fluency** than unmasked ones (Wave 1: $\Delta R^2 = 0.044$; Wave 2: $\Delta R^2 = 0.003$). These effects were much more pronounced than in prior analyses with over 20% of the variance in Wave 2 fluency uniquely associated with masked performance. This finding again suggests that automaticity, measured by masked performance, is critical for fluency, and this may be particularly true for multisyllabic words among adolescents. Unmasked performance accounted for unique variance in **decoding** scores ($\Delta R^2 = 0.112$) in Wave 1, whereas masked performance did not ($\Delta R^2 = 0.009$). In Wave 2, however, both unmasked ($\Delta R^2 = 0.023$) and masked task versions ($\Delta R^2 = 0.095$) contributed significant variance, and the ΔR^2 was higher for masked performance than unmasked one. For **comprehension** scores, masked performance significantly contributed unique variance in both waves (Wave 1: $\Delta R^2 = 0.063$; Wave 2: $\Delta R^2 = 0.084$), whereas unmasked performance did not (Wave 1: $\Delta R^2 = 0.002$; Wave 2: $\Delta R^2 = 0.007$).

Although the pattern of relative effect sizes was similar to the earlier analyses of monosyllabic words, the unique variance associated tended to be greater for multisyllabic words (though the overall R^2s were similar). This suggests that multisyllabic words may increase the power to detect the unique contributions of automaticity to fluency and comprehension.

Analysis 4: Relative Contributions of Individual Tasks

To assess contributions of different tasks (Question 4), we asked which tasks individually reached significance on Step 3 when all factors were entered. We concentrated on fluency and decoding, as neither unmasked nor masked task/versions consistently predicted comprehension. All analyses can be found in the online supplement (S1). To summarize the main findings, for fluency there was not one masked task that was uniquely predictive. However, for decoding, unmasked versions of *Find the Rhyme* and *Verify* (when *Find the Rhyme* was not included in the analyses of multisyllabic tasks) account for most of the variance. This makes sense as these tasks were predicted to most strongly tap the O→P pathways required for decoding.

Analysis 5: Automaticity as a Stable Trait

Finally, we asked whether a participants' degree of automaticity was stable (Question 5). To answer this, we needed a measure of automaticity (masked performance) that was statistically independent of knowledge (unmasked). Thus, we first predicted performance on masked trials of a task from unmasked performance in the same task and oral vocabulary (PPVT). We then computed whether subjects performed better or worse on the masked trials than predicted from unmasked. This was done separately at each testing point in Wave 2.

This analysis was implemented in a mixed effects model in R (version Ri386 3.5.3; R Core Team, 2013). In this model, accuracy on the masked trials was the dependent variable. Accuracy on each unmasked task version (*Find the Picture*, *Find the Rhyme*, *Ve*rify, *Find the Picture* [multisyllabic], and *Verify* [multisyllabic]) was computed separately for each subject, and used to predict masked performance (on that same task). Thus, the model included fixed factors of task (dummy coded), performance on the unmasked version of that task, the task × unmasked interaction and oral vocabulary. This model was fit separately for data from the initial collection and the reliability collection for the subset of Wave 2 in the reliability study ($N = 122$).

Subsequently, the random intercept of each subject was extracted as an estimate of that subject's ability on the masked tasks after accounting for these other factors. The rationale behind this analysis is as follows: If automaticity behaves as a stable trait, its relationship to participants' unmasked performance should be reliable across different testing points. Consequently, participants' subject intercepts (their latent ability) from each testing should be correlated. The results showed a strong correlation between subject slopes at each time point, $r = 0.75$ ($t(120) = 12.28$, $p < 0.001$), indicating that participants' automaticity was individually reliable across them (Figure 3), even after accounting for knowledge of the relevant tasks and items.

Discussion

We discuss the results with respect to each of the research questions. We then highlight some limitations before concluding.

Contributions of Automaticity and Knowledge to Reading Outcomes (Questions 1, 5)

We first investigated the contributions of automaticity and knowledge in monosyllabic words to decoding, fluency, and reading comprehension in our sample of average to below-average middle-schoolers. Performance on all tasks accounted for a large amount of variance in reading outcomes. This was particularly surprising for fluency because none of the tasks asked for an oral response, and all used single words. The shared variance was always large: This was true for all reading outcomes, including ones that are consistently better predicted by one task version than the other (see also Roembke et al., 2019).

When we turned to unique variance, in general, masked performance uniquely predicted fluency, whereas unmasked performance predicted decoding. These results are consistent with Lovett (1987) and Roembke et al. (2019), and support theoretical accounts that predict automaticity's importance for fluency (LaBerge & Samuels, 1974; Perfetti, 1985). Our study significantly extends prior work (e.g., Cutting & Scarborough, 2006), by using a new measure which is robust to processing speed, differences in the speed/accuracy trade-off, and which accounts for the general level of knowledge (via the unmasked task versions).Thus, automaticity—which is well-captured by masked performance—is a unique predictor of fluency that is independent of overall task performance and relevant reading knowledge.

In contrast, knowledge of letters, GPC regularities, and sight words—measured by unmasked performance—accounted for unique variance in decoding (whereas masked performance did not). This is reasonable given that standardized tests of decoding and word recognition (as were used here) are typically untimed. The role of automaticity is less clear for reading comprehension (though see discussion of multisyllabic tasks, Question 3).

The observed double dissociation between knowledge and automaticity further supports the notion of two independent traits—knowledge and automaticity—that are critical for distinct outcomes (Lovett, 1987; Roembke et al., 2019; Stanovich, 1980). We note that these traits cannot be truly orthogonal—a child cannot be an automatic word recognizer if they lack the knowledge to read words, and developmentally the contribution of automaticity and knowledge to fluency may be complex. However, our work shows that each plays a unique role in outcomes. We further investigated this by asking whether the contributions of automaticity—the main trait of interest—to reading outcomes remained individually stable across testing points (Question 5). This was indeed the case, as the correlation of automaticity estimates from the two testing points was high ($r = 0.75$).

The strength of the correlation adds to the evidence that automaticity (measured by masked performance) may serve as a stable marker of students' skill profile that should be accounted for when assessing why a student may struggle with reading. This finding is consistent with previous studies, showing that RAN remains a stable predictor of fluency throughout reading development (see Protopapas, Altani, & Georgiou, 2013 for an overview) and that fluency was stable across eight years (Landerl & Wimmer, 2008). Critically, the reliability estimate reported here was observed after accounting for unmasked performance. A student who fails when items are masked could fail because they are not automatic enough or because they simply cannot read those words. That is, masked performance is not meaningful without knowing the student's capacity to do the task without masking. Despite accounting for this—removing a good portion of the relevant variance across children—we observed large effects.

Automaticity in Nonword Processing (Question 2)

Question 2 asked if automaticity in word-level processes results from the direct activation of meanings from written words (e.g., the O→S pathway) or if mappings from orthography to phonology (O→P) can also be automatic. We addressed this by asking whether automaticity in nonword processing accounts for unique variance in outcomes. Our initial analyses of accuracy suggest that the O→P pathway functions much like the O→S pathway. There were few interactions between masking and item-type on accuracy. A uniform effect of masking may suggest that automatic recognition of nonwords relied on similar processes as word processing.

Importantly, our communality approach showed that automaticity predicts unique variance in fluency. This is not likely due to differences in decoding, as we controlled for decoding knowledge in the first step of the analysis and nonwords were counterbalanced between masked and unmasked conditions. These analyses confirm that automaticity can exist within O→P mappings, and that differences in how quickly or automatically one can activate a nonword can predict fluency—even fluency measured with real sentences. Further, the fact that variation within a pathway was meaningful suggests that automaticity is not solely achieved by switching to a more direct pathway (e.g., memorizing the words). Rather, there can be variation in the automaticity of individual pathways. Here, we document this unambiguously in the O→P mappings, but there is no reason to assume this cannot be observed in O→S.

Thus, the ability to process novel letter strings automatically might be an important skill when reading, and its absence might contribute to reading difficulties. The similarity between the O→S and the O→P pathway suggests that automaticity within both pathways may develop in parallel. Future longitudinal work should explore the developmental timeline of the two pathways as well as the relative importance of word and nonword contributions, asking whether performance on nonword trials accounts for variance over and above of word trials.

Automaticity of Multisyllabic Words (Question 3)

We found that the pattern of linkage between automatic (masked) and non-automatic (unmasked) processing and reading outcomes remained the same for multisyllabic words as with monosyllabic ones (Question 3): masked performance uniquely predicted fluency, whereas unmasked performance uniquely predicted decoding. Particularly for fluency, multisyllabic items showed substantially larger effects (upwards of 20% in Wave 2) and were potentially more diagnostic than monosyllabic items. Multisyllabic items also showed a contribution of automaticity to comprehension which was not evident with monosyllabic words. Comprehension plays a central role both practically and theoretically (LaBerge & Samuels, 1974; Perfetti, 1985) as the ultimate goal of reading instruction. This offers the first evidence that automaticity is a critical predictor of reading comprehension.

The inclusion of multisyllabic words may be particularly important for applying paradigms like this to adolescents of a wider range of abilities (i.e., those who are above-average readers), as our sample consisted of average and below-average readers. This stands in contrast to measures like RAN that require simple, well-known words (e.g., Denckla & Rudel, 1974; Lervåg & Hulme, 2009). Use of masking may enable a test of automaticity in challenging words, which are far more age appropriate for struggling middle-school readers, and which may be more important for students as they progress to more complex text. Of course, classroom assessments that build on our research would have to be shortened (e.g., by additional Item Response Theory analyses) and refined, but our experimental paradigm offers helpful data points for such avenues.

In addition, these findings suggest that difficulty in reading multisyllabic words is at least partly due to a lack of automaticity: Below-average readers may struggle with multisyllabic words not only because they are harder to decode or recognize (even if given unlimited time), but also because they are harder to activate automatically. Developing automaticity of multisyllabic words might be slowed down due to several reasons (e.g., number of letters, lower frequency in text), and future research should explore this in more detail.

Unique Contributions of Automaticity to Outcomes across Tasks (Question 4)

Finally, we explored whether individual tasks differentially predicted outcomes. *Find the Picture* and *Find the Rhyme* were intended to maximize involvement of distinct reading pathways (O→S and O→P respectively, even as both pathways are likely involved in all tasks). The O→S pathway, as it is more direct and thus quicker, might be more important for automaticity (e.g., Harm & Seidenberg, 2004). This was hinted at by our previous work (Roembke et al., 2019) where *Find the Picture* accounted for the most variance in fluency. In contrast, *Find the Rhyme* might be more important when unmasked in predicting decoding (see Table 1).

The unique contributions of each task were not stable across waves and no single masked task consistently predicted fluency. Instead, automaticity appears to be well-captured by a range of tasks, suggesting that words can be accessed automatically via different pathways when necessary. This reinforces the utility of testing constructs like automaticity (or knowledge) across multiple tasks to avoid practice effects and engage students' interest. There was a tendency for the unmasked version of *Find the Rhyme* to account for most variance in decoding. This suggests that students' knowledge is best reflected in tasks that stress words' phonological components, whereas automaticity might be captured with a range of tasks.

The Development of Automaticity of Word-Level Processes

Our results suggest that automaticity exists both in the direct and indirect reading pathways, consistent with thinking in many domains of learning (see Moors & De Houwer, 2006). Mechanistically, this likely derives as a form of practice: As a result of exposure to written words and/or repeated reading practice, activation may spread more easily from orthographic representations to both phonological and semantic ones (e.g., Plaut, McClelland, & Seidenberg, 2014; Seidenberg, 2005). This is consistent with how changes in reading speed and comprehension happen at the sentence level (Wells, Christiansen, Race, Acheson, & MacDonald, 2009), where exposure to less frequent relative clause structures has led to increased reading speed.

However, given the cross-sectional design of this study, we can only speculate on how automaticity *develops* in each pathway. We have mostly considered a unidirectional model based on the Simple View of Reading (Gough & Tunmer, 1986) in which increases in automaticity lead to changes in fluency that in turn frees resources, enabling children to comprehend more in the moment (see also LaBerge & Samuels, 1974; Perfetti, 1985). This is not a developmental model in the sense that automaticity does not lead to later changes in comprehension. Rather, the idea is that automaticity shapes comprehension in the moment. However, a developmental model along these lines is possible: As children recognize words more automatically, they find reading easier, are more likely to read more, and therefore become more automatic via practice.

However, the relationship between automaticity and fluency is likely complex where changes in fluency can also influence the development of automaticity. A key factor in the development of automaticity more broadly is practice (Anderson, 1992; Logan, 1985, 1997; Palmeri, 1999). It seems likely that increases in comprehension could facilitate recognition of words in their oral vocabulary using context. This could in turn enable children to get greater practice benefits from reading, and thus to facilitate later automatic processing. Similarly, as decoding becomes more automatic, children may be in a better position to rapidly read words they have not seen before, enabling the encoding of direct O→S mappings.

Future research should explore the exact developmental timeline of how these different constructs relate to each other. Here, masking may be an effective tool for isolating automaticity in longitudinal contexts. Ultimately, however, all these models suggest that the automaticity of word recognition could play multiple roles in reading development.

Intervention studies could help disentangle these by targeting the development of automaticity in specific pathways.

Limitations

This study builds on the previously conducted analyses by Roembke et al. (2019) by more than quadrupling the sample size per wave, extending the set of items, and assessing stability. However, there are several limitations. First, overall accuracy was high. This was particularly true in *Find the Picture* and *Verify*. As a result, reliability was lower than if performance had shown more variability (e.g., *Find the Picture/unmasked* had the lowest Cronbach's alpha but highest accuracy). Nevertheless, alphas were respectable for all tasks, and we observed high reliability of automaticity as a trait when estimated across all tasks. To investigate the potential impact of ceiling effects, we replicated the monosyllabic analyses for Wave 1 without *Find the Picture* and the pattern of results remained the same (see online supplement S2).

Second, results across waves were not always consistent, particularly for comprehension. These inconsistencies may derive from variability in the student populations (e.g., slight age differences). However, they may also derive from differences in the experimental paradigm across waves, based on an IRT analysis after Wave 1 that identified the best items and shortened the experiment. Thus, results from Wave 2 may include less measurement noise than Wave 1. Importantly, the relative magnitude of effects remained relatively stable across waves for decoding and fluency, though less so for comprehension. Thus, even as results from a single sample should not be over-interpreted, focusing on effect sizes may offer more clarity.

Third, similar to other cognitive tasks, the backward masking tasks we employed here likely taps other cognitive skills besides automaticity (e.g., attention, ability to hold information in working memory and motivation). However, one advantage of our design remains that some of these task demands can be captured by the unmasked versions of the same tasks. It may be argued that masked task versions tap these domain general skills to a greater extent than the unmasked ones. However, if these skills were driving the effects, unique variance of masked performance should have also predicted comprehension and decoding. This was not observed. Similarly, while backwards masking may eliminate the contribution of speed of processing *outside* the reading system, components attributed to word reading speed may also relate to general speed of processing (which would also be relevant for fluency). The present data cannot disentangle these contributions. However, at a functional level, our masking paradigm can isolate how fast children read words from differences in general cognitive processing, and this shows clearly that specific word reading speed is crucial for fluency.

Fourth, even as our results underscore automaticity as a complementary trait to knowledge in the specific sample of average to struggling middle-school readers tested, it is not clear whether this finding generalizes to other languages than English. English—with its high number of quasi-regularities and exception words (Seidenberg, 2005)—may create unique pressures on readers to learn direct O→S mappings, whereas readers of more transparent languages can simply decode words. Alternatively, there could be less variability in readers of transparent languages, thus reducing automaticity's importance in predicting outcomes. However, supporting the converse, some studies find RAN to be more predictive of reading in transparent than in opaque languages (e.g., Wimmer, Mayringer, & Landerl, 2000, but see also Ziegler et al., 2010).

Finally, the fact that our sample was restricted to average-to-below-average readers may constrain generalizability. The most likely consequence of this restricted range would be to depress the correlations—particularly on comprehension, which was the basis of the restriction.

This is because less variation on the predictors makes it harder to obverse a correlation. Consequently, the true correlations may be higher than what was observed here. However, it is also possible that the contributions of automaticity to reading outcomes differs at different levels of ability. Just as with transparent languages, however, higher-achieving readers could show a reduced or an increased contribution of automaticity. Future research with a full range of middle-school students (preferably in students with different native languages) should be conducted to clarify the generalizability of our results, and quantile regression may be helpful for asking if predictive relationships hold at all levels. Nevertheless, it is important to note that the ability range studied here is perhaps most important for understanding the role of automaticity, as this is the range where instructional design choices and interventions are likely to be most critical.

Conclusions

Overall, these results suggest that average to below-average middle-school students' word recognition can be captured by two independent, stable traits: knowledge and automaticity. Automaticity was more critical in accounting for average to struggling students' fluency and potentially comprehension (particularly in the context of multisyllabic words), whereas knowledge predicted decoding. These findings indicate that a lack in automaticity in word-level processes is a plausible bottleneck in middle-school students' reading. These results further support masking as an assessment of automaticity in word-level reading skills, particularly in combination with students' unmasked performance. This adds to a growing appreciation of the independent contributions of automatic (rapid) processing and knowledge to development and disorders in a variety of domains including oral language development (McMurray, Danelz, Rigler, & Seedorff, 2018; McMurray, Horst, & Samuelson, 2012), mathematics (e.g., Cumming & Elkins, 1999) and even social cognition (Bargh, Schwader, Hailey, Dyer, & Boothby, 2012).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

Anderson JR (1992). Automaticity and the ACT Theory. The American Journal of Psychology, 105(2), 165–180. 10.2307/1423026 [PubMed: 1621879]

Bargh JA, Schwader KL, Hailey SE, Dyer RL, & Boothby EJ (2012). Automaticity in social-cognitive processes. Trends in Cognitive Sciences, 16(12), 593–605. 10.1016/j.tics.2012.10.002 [PubMed: 23127330]

Bruck M. (1990). Word-recognition skills of adults with childhood diagnoses of dyslexia. Developmental Psychology, 26(3), 439–454. 10.1037/0012-1649.26.3.439

Catts HW, Fey ME, Zhang X, & Tomblin JB (1999). Language basis of reading and reading disabilities: evidence from a longitudinal investigation. Scientific Studies of Reading, 3(4), 331–361. 10.1207/s1532799xssr0304_2

Cirino PT, Romain MA, Barth AE, Tolar TD, Fletcher JM, & Vaughn S. (2013). Reading skill components and impairments in middle school struggling readers. Reading and Writing, 26(7), 1059–1086. 10.1007/s11145-012-9406-3 [PubMed: 24000271]

Coltheart M, Rastle K, Perry C, Langdon R, & Ziegler JC (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. Psychological Review, 108(1), 204–256. 10.1037/0033-295X.108.1.204 [PubMed: 11212628]

Cumming J, & Elkins J. (1999). Lack of automaticity in the basic addition facts as a characteristic of arithmetic learning problems and instructional needs. Mathematical Cognition, 5(2), 149–180. 10.1080/135467999387289

Cutting LE, & Scarborough HS (2006). Prediction of reading comprehension: relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. Scientific Studies of Reading, 10(3), 277–299. 10.1207/s1532799xssr1003_5

De Jong PF, & Van Der Leij A. (1999). Specific contributions of phonological abilities to early reading acquisition: Results from a Dutch latent variable longitudinal study. Journal of Educational Psychology, 91(3), 450–476. 10.1037/0022-0663.91.3.450

Denckla MB, & Rudel RG (1974). Rapid "automatized" naming of pictured objects, colors, letters and numbers by normal children. Cortex, 10(2), 186–202. 10.1016/S0010-9452(74)80009-2 [PubMed: 4844470]

Dickman SJ, & Meyer DE (1988). Impulsivity and speed-accuracy tradeoffs in information processing. Journal of Personality and Social Psychology, 54(2), 274–290. 10.1037/0022-3514.54.2.274 [PubMed: 3346814]

Ferguson CA, & Farwell CB (1975). Words and sounds in early language acquisition: English initial consonants in the first fifty words. Language, 51(2), 419–439. 10.2307/412864

Forster KI, & Davis CJ (1984). Repetition priming and frequency attenuation in lexical access. Journal of Experimental Psychology: Learning, Memory, and Cognition, 10(4), 680–698. 10.1037/0278-7393.10.4.680

Francis DJ, Barth AE, Cirino PT, Reed DK, & Fletcher JM (2010). Texas middle school fluency assessment, version 2.0. Houston, TX: University of Houston/Texas Education Agency.

Gough PB, & Tunmer WE (1986). Decoding, reading, and reading disability. Remedial and Special Education, 7(1), 6–10. 10.1177/074193258600700104

Harm MW, & Seidenberg MS (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. Psychological Review, 111(3), 662–720. 10.1037/0033-295X.111.3.662 [PubMed: 15250780]

Jared D, & Seidenberg MS (1990). Naming multisyllabic words. Journal of Experimental Psychology: Human Perception and Performance, 16(1), 92–105. 10.1037/0096-1523.16.1.92 [PubMed: 2137526]

Jenkins JR, Fuchs LS, van den Broek P, Espin CA, & Deno SL (2003). Sources of individual differences in reading comprehension and reading fluency. Journal of Educational Psychology, 95(4), 719–729. 10.1037/0022-0663.95.4.719

Kail R, & Hall LK (1994). Processing speed, naming speed, and reading. Developmental Psychology, 30(6), 949–954. 10.1037/0012-1649.30.6.949

Klauda SL, & Guthrie JT (2008). Relationships of three components of reading fluency to reading comprehension. Journal of Educational Psychology, 100(2), 310–321. 10.1037/0022-0663.100.2.310

LaBerge D, & Samuels SJ (1974). Toward a theory of automatic information processing in reading. Cognitive Psychology, 6(2), 293–323. 10.1016/0010-0285(74)90015-2

Landerl K, & Wimmer H. (2008). Development of word reading fluency and spelling in a consistent orthography: An 8-year follow-up. Journal of Educational Psychology, 100(1), 150–161. 10.1037/0022-0663.100.1.150

Lervåg A, & Hulme C. (2009). Rapid Automatized Naming (RAN) taps a mechanism that places constraints on the development of early reading fluency. Psychological Science, 20(8), 1040–1048. 10.1111/j.1467-9280.2009.02405.x [PubMed: 19619178]

Logan GD (1985). Skill and automaticity: Relations, implications, and future directions. Canadian Journal of Psychology/Revue Canadienne de Psychologie, 39(2), 367–386. 10.1037/h0080066

Logan GD (1997). Automaticity and reading: perspectives from the instance theory of automatization. Reading & Writing Quarterly, 13(2), 123–146. 10.1080/1057356970130203

Lovett MW (1987). A developmental approach to reading disability: Accuracy and speed criteria of normal and deficient reading skill. Child Development, 58(1), 234–260. 10.1111/j.1467-8624.1987.tb03503.x [PubMed: 3816346]

MacGinitie WH, MacGinitie RK, Maria K, & Dreyer L. (2000). Gates-MacGinitie reading tests, 4th edition. Chicago, IL: Riverside Publishing.

Manis FR (1985). Acquisition of word identification skills in normal and disabled readers. Journal of Educational Psychology, 77(1), 78–90. 10.1037/0022-0663.77.1.78

McMurray B, Danelz A, Rigler H, & Seedorff M. (2018). Speech categorization develops slowly through adolescence. Developmental Psychology, 54(8), 1472–1491. 10.1037/dev0000542 [PubMed: 29952600]

McMurray B, Horst JS, & Samuelson LK (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. Psychological Review, 119(4), 831–877. 10.1037/a0029872 [PubMed: 23088341]

Moors A, & De Houwer J. (2006). Automaticity: A theoretical and conceptual analysis. Psychological Bulletin, 132, 297–326. 10.1037/0033-2909.132.2.297 [PubMed: 16536645]

Morris D, Trathen W, Frye EM, Kucan L, Ward D, Schlagal R, & Hendrix M. (2013). The role of reading rate in the informal assessment of reading ability. Literacy Research and Instruction, 52, 52–64. 10.1080/19388071.2012.702188

Morris D, Trathen W, Lomax RG, Perney J, Kucan L, Frye EM, … Schlagal R. (2012). Modeling aspects of print-processing skill: Implications for reading assessment. Reading and Writing: An Interdisciplinary Journal, 25, 189–215. 10.1007/s11145-010-9253-z

National Institute of Child Health and Human Development. (2000). Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction. In NIH Publication 004754. 10.1002/ppul.1950070418

Nippold MA (2017). Reading comprehension deficits in adolescents: Addressing underlying language abilities. Language, Speech, and Hearing Services in Schools, 48, 125–131. 10.1044/2016_lshss-16-0048

Oslund EL, Clemens NH, Simmons DC, Smith SL, & Simmons LE (2016). How vocabulary knowledge of middle-school students from low socioeconomic backgrounds influences comprehension processes and outcomes. Learning and Individual Differences, 45, 159–165. 10.1016/j.lindif.2015.11.013

Palmeri TJ (1999). Theories of automaticity and the power law of practice. Journal of Experimental Psychology: Learning Memory and Cognition, 25(2), 543–551. 10.1037/0278-7393.25.2.543

Perfetti CA (1985). Reading ability. New York: Oxford University Press.

Plaut DC, McClelland JL, & Seidenberg MS (2014). Reading exception words and pseudo words: Are two routes really necessary? In Connectionist Models of Memory and Language. 10.4324/9781315794495

Protopapas A, Altani A, & Georgiou GK (2013). Development of serial processing in reading and rapid naming. Journal of Experimental Child Psychology, 116(4), 914–929. 10.1016/j.jecp.2013.08.004 [PubMed: 24077466]

Rasinski TV, Padak ND, McKeon CA, Wilfong LG, Friedauer JA, & Heim P. (2005). Is reading fluency a key for successful high school reading? Journal of Adolescent and Adult Literacy, 49(September), 22–27. 10.1598/JAAL.49.1.3

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Reed DK, Martin E, Hazeltine E, & McMurray B. (2019). Students' perceptions of a gamified reading assessment. Journal of Special Education Technology, 35(4), 191–203. 10.1177/0162643419856272

Reicher GM (1969). Perceptual recognition as a function of meaningfulness of stimulus material. Journal of Experimental Psychology, 81(2), 275–280. 10.1037/h0027768 [PubMed: 5811803]

Roembke TC, Hazeltine E, Reed DK, & McMurray B. (2019). Automaticity of word recognition is a unique predictor of reading fluency in middle-school students. Journal of Educational Psychology, 111(2), 314–330. 10.1037/edu0000279

Seidenberg MS (2005). Connectionist models of word reading. Current Directions in Psychological Science, 14(5), 238–242. 10.1111/j.0963-7214.2005.00372.x

Seidenberg MS, & McClelland JL (1989). A distributed, developmental model of word recognition and naming. Psychological Review, 96(4), 523–568. 10.1037/0033-295X.96.4.523 [PubMed: 2798649]

Share DL (1995). Phonological recoding and self-teaching: sine qua non of reading acquisition. Cognition, 55(2), 151–218. 10.1016/0010-0277(94)00645-2 [PubMed: 7789090]

Share DL (2008). Orthographic learning, phonological recoding, and self-teaching. Advances in Child Development and Behavior, 36, 31–82. 10.1016/S0065-2407(08)00002-5 [PubMed: 18808041]

Sorensen LC (2019). "Big Data" in educational administration: An application for predicting school dropout risk. Educational Administration Quarterly, 55, 404–446. 10.1177/0013161X18799439

Stanovich KE (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. Reading Research Quarterly, 16(1), 32–71. 10.2307/747348

Tamura N, Castles A, & Nation K. (2017). Orthographic learning, fast and slow: Lexical competition effects reveal the time course of word learning in developing readers. Cognition, 163, 93–102. 10.1016/j.cognition.2017.03.002 [PubMed: 28314178]

Torgesen JK, & Hudson RF (2006). Reading fluency: Critical issues for struggling readers. In Samuels SJ & Farstrup AE (Eds.), What research has to say about fluency instruction (pp. 130–158). Newark, DE, US: International Reading Association.

Walczyk JJ (2000). The interplay between automatic and control processes in reading. Reading Research Quarterly, 35(4), 554–566. 10.1598/rrq.35.4.7

Wells JB, Christiansen MH, Race DS, Acheson DJ, & MacDonald MC (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. Cognitive Psychology, 58(2), 250–271. 10.1016/j.cogpsych.2008.08.002 [PubMed: 18922516]

Wheeler DD (1970). Processes in word recognition. Cognitive Psychology, 1(1), 59–85. 10.1016/0010-0285(70)90005-8

Wimmer H, Mayringer H, & Landerl K. (2000). The double-deficit hypothesis and difficulties in learning to read a regular orthography. Journal of Educational Psychology, 92(4), 668–680. 10.1037/0022-0663.92.4.668

Woodcock RW, McGrew K, & Mather N. (2001). Woodcock-Johnson Tests of Achievement (3rd ed.). Itasca, IL: Riverside.

Ziegler JC, Bertrand D, Tóth D, Csépe V, Reis A, Faísca L, … Blomert L. (2010). Orthographic depth and its impact on universal predictors of reading: A cross-language investigation. Psychological Science, 21(4), 551–559. 10.1177/0956797610363406 [PubMed: 20424101]

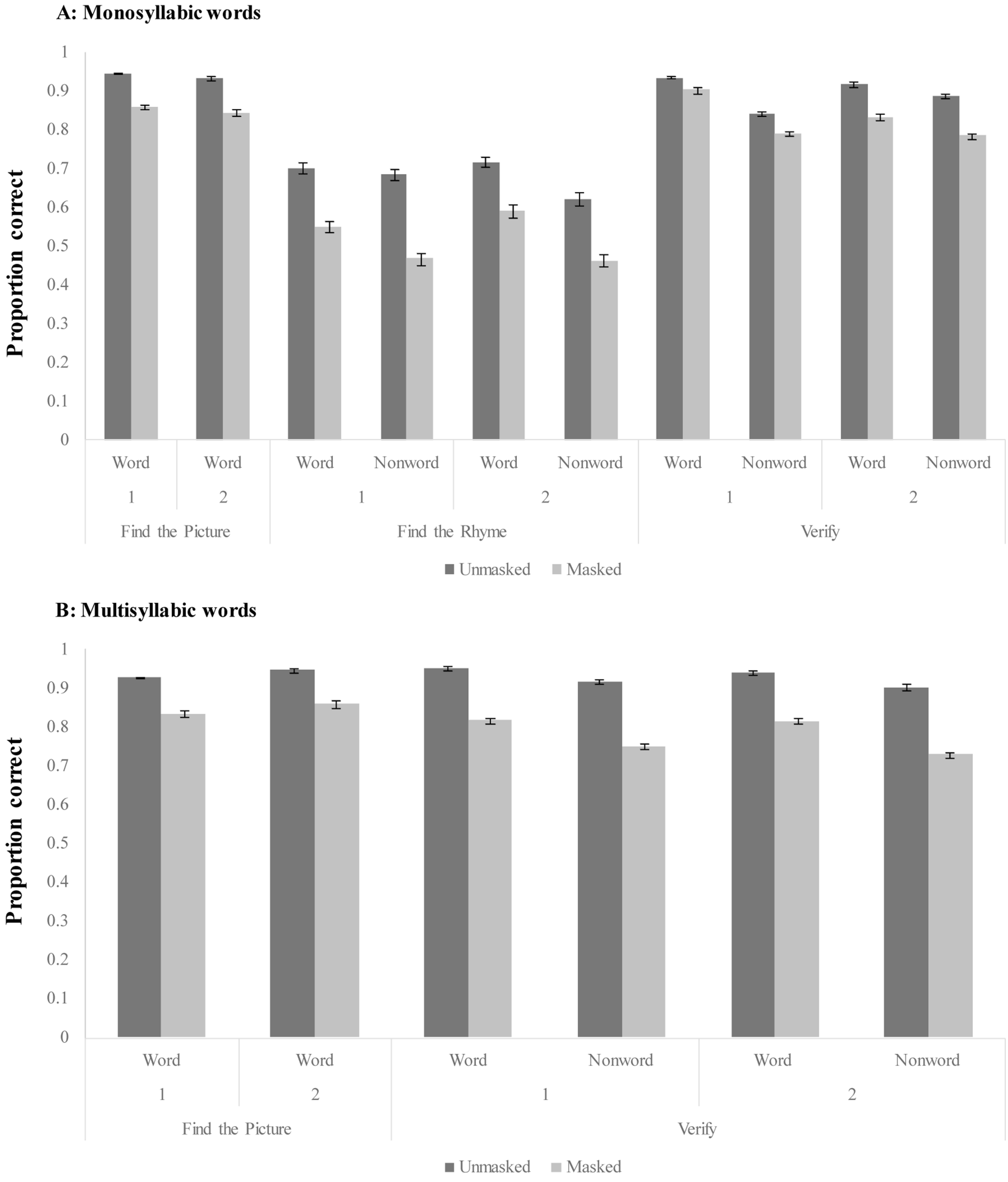

Figure 1.

Overview of performance on different task versions. Performance is plotted as raw accuracy, not transformed, to facilitate readability. The numerals one and two refer to the two waves. Panel A includes only monosyllabic tasks; Panel B presents multisyllabic tasks. Error bars indicate standard error of the mean.

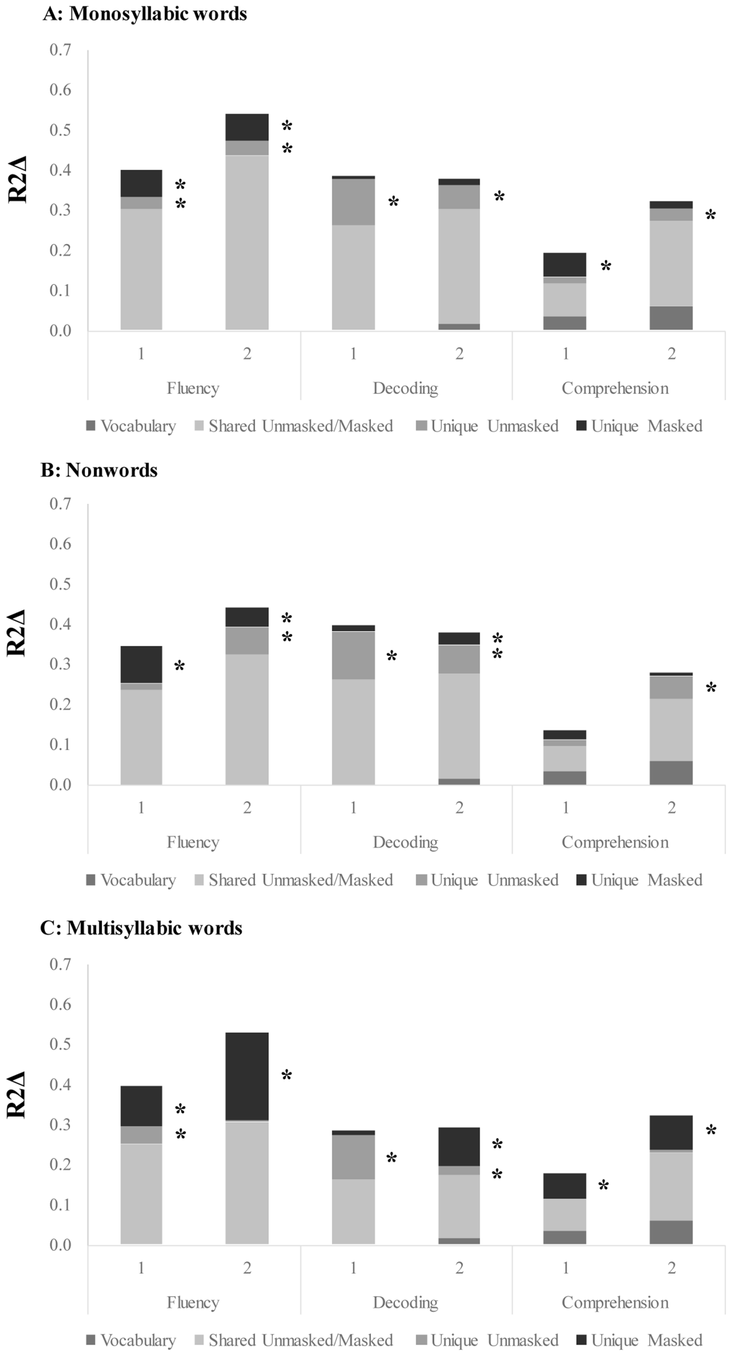

Figure 2.

Unique and shared variance associated with masked and unmasked performance for Wave 1 and Wave 2 for monosyllabic tasks (Panel 1), nonword trials (Panel B) and multisyllabic tasks (Panel C). * indicates unique variance reaching significance.

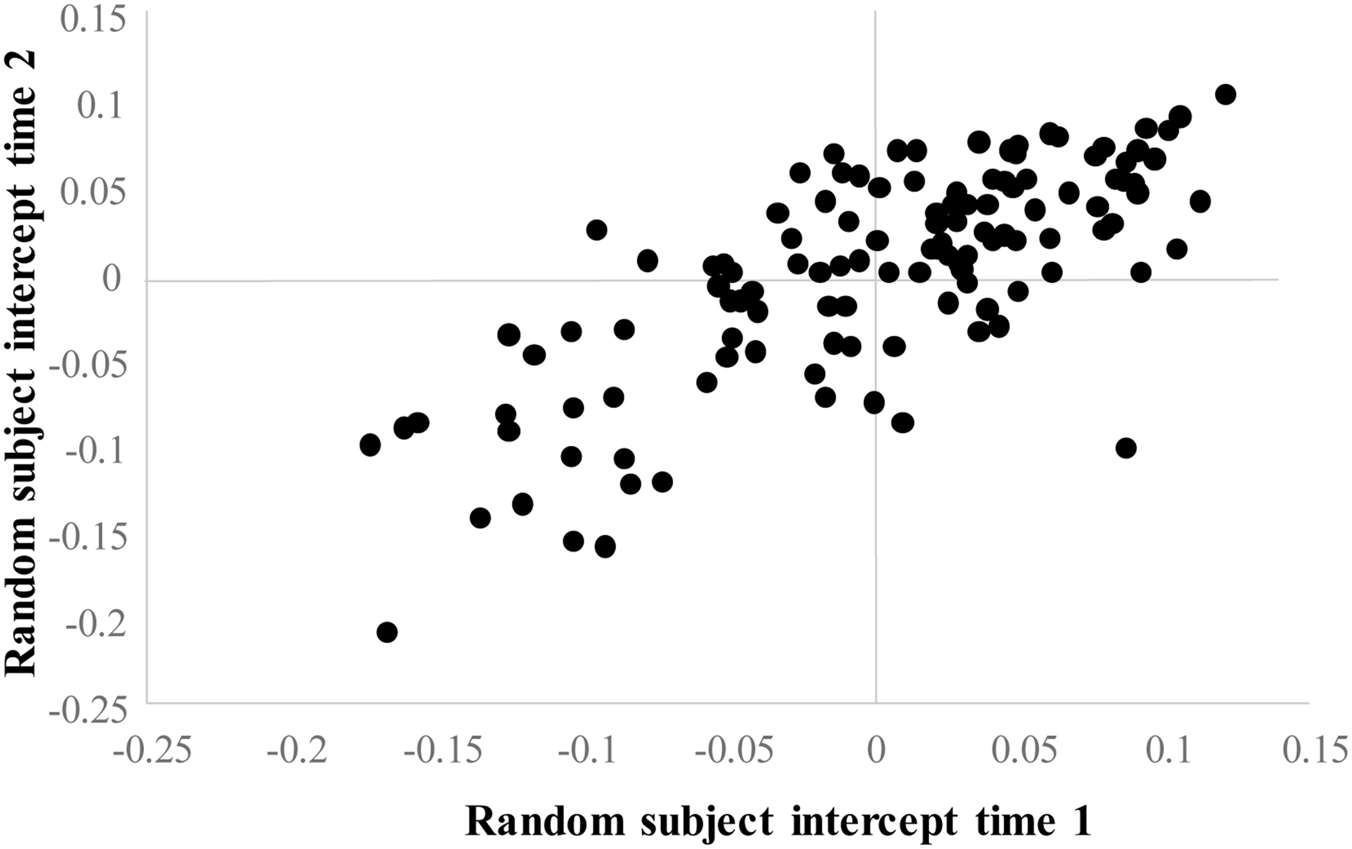

Figure 3.

Random subject intercepts for masked tasks at time point 1 (initial testing) and time point 2 (reliability testing).

Table 1

Overview of tasks. All tasks existed in unmasked and masked versions. Multisyllabic word versions were not included in Roembke et al. (2019).

Task name	Description	Nonword version?	Pathway targeted?	Multisyllabic word version?
Find the Picture	Participants see a written target word and four visual representations. They select the picture that matches the target word.	No	O→S	Yes
Find the Rhyme	Participants see a written target word and eight written response options. They select the response option that rhymes with the target word.	Yes	O→P→S	No
Verify	Participants hear a word and see a written target word. They indicate whether the two words matched or not.	Yes	O→P→S; O→S	Yes

Table 2

Overview of reading outcomes.

Variable	Wave 1	Wave 2	Wave 2 Reliability Subset
	Mean ± SD	Mean ± SD	Mean ± SD
WRMT	90.23 ± 13.14	91.61 ± 13.18	92.16 ± 13.26
TMSFA	137.27 ± 27.59	118.44 ± 33.61	117.55 ± 31.93
GMRT	92.25 ± 9.00	94.86 ± 10.08	94.61 ± 9.96
PPVT	96.22 ± 9.02	96.05 ± 9.74	96.74 ± 10.11

Note. All measures are given as standardized scores except for TMSFA, which is reported as the average equated value of the number of words read correctly per minute. The WRMT is the average score of participants' WRMT_ID (Word identification) and WRMT_AT (word attack) score. WRMT = Woodcock Reading Mastery; Texas Middle School Fluency Assessment; GMRT = Gates-MacGinitie Reading Test; PPVT = Peabody Picture Vocabulary Test.

Table 3

Comparative analyses for masked and unmasked versions of each task. Effect of item-type (and interaction) not shown for the Find the Picture Task, as nonwords were not possible in that task so these effects were not compared. (Picture = Find the Picture; Rhyme = Find the Rhyme; Mono = monosyllabic words; Multi = multisyllabic words) for both waves.

Word Length	Task	Wave	Masking	Item-Type	Masking × Item-Type
Mono	Picture	1	$F(1, 229) = 335.26$, $p < 0.001$, $\eta_P^2 = 0.594$	—	—
		2	$F(1, 213) = 187.51$, $p < 0.001$, $\eta_P^2 = 0.468$	—	—
	Rhyme	1	$F(1, 229) = 489.82$, $p < 0.001$, $\eta_P^2 = 0.681$	$F(1, 229) = 40.12$, $p < 0.001$, $\eta_P^2 = 0.149$	$F(1, 229) = 42.74$, $p < 0.001$, $\eta_P^2 = 0.157$
		2	$F(1, 213) = 208.29$, $p < 0.001$, $\eta_P^2 = 0.494$	$F(1, 213) = 215.06$, $p < 0.001$, $\eta_P^2 = 0.502$	$F(1, 213) = 1.23$, $p = 0.269$, $\eta_P^2 = 0.006$
	Verify	1	$F(1, 229) = 362.61$, $p < 0.001$, $\eta_P^2 = 0.613$	$F(1, 229) = 93.39$, $p < 0.001$, $\eta_P^2 = 0.290$	$F(1, 229) = 0.04$, $p = 0.843$, $\eta_P^2 < 0.001$
		2	$F(1, 213) = 271.30$, $p < 0.001$, $\eta_P^2 = 0.560$	$F(1, 213) = 77.90$, $p < 0.001$, $\eta_P^2 = 0.268$	$F(1, 213) = 0.01$, $p = 0.915$, $\eta_P^2 < 0.001$
Multi	Picture	1	$F(1, 229) = 303.36$, $p < 0.001$, $\eta_P^2 = 0.570$	—	—
		2	$F(1, 213) = 292.57$, $p < 0.001$, $\eta_P^2 = 0.579$	—	—
	Verify	1	$F(1, 229) = 853.75$, $p < 0.001$, $\eta_P^2 = 0.789$	$F(1, 229) = 123.60$, $p < 0.001$, $\eta_P^2 = 0.351$	$F(1, 229) = 0.04$, $p = 0.852$, $\eta_P^2 < 0.001$
		2	$F(1, 213) = 569.48$, $p < 0.001$, $\eta_P^2 = 0.728$	$F(1, 213) =170.77$, $p < 0.001$, $\eta_P^2 = 0.445$	$F(1, 213) = 2.09$, $p = 0.150$, $\eta_P^2 = 0.010$

Table 4

Overview of communality analysis including shared and unique variance for masked and unmasked tasks for monosyllabic tasks, nonword trials of masked and unmasked monosyllabic tasks and multisyllabic tasks.

			Overall	Shared	Unmasked	Masked
	Outcome	**Wave**	***R*** ^{2}	**Δ*R*** ^{2}	**Δ*R*** ^{2}	**Δ*R*** ^{2}
Monosyllabic	Fluency	1	0.399	0.301	0.030*	0.065**
		2	0.539	0.435	0.037*	0.067**
	Decoding	1	0.383	0.262	0.112**	0.008
		2	0.379	0.289	0.060**	0.015
	Comprehension	1	0.194	0.084	0.015	0.059*
		2	0.321	0.211	0.030*	0.019
Nonwords	Fluency	1	0.346	0.234	0.016	0.093**
		2	0.439	0.327	0.064**	0.049**
	Decoding	1	0.397	0.262	0.119**	0.014
		2	0.376	0.263	0.069**	0.030*
	Comprehension	1	0.132	0.060	0.017	0.020
		2	0.281	0.154	0.054*	0.010
Multisyllabic	Fluency	1	0.397	0.250	0.044**	0.100**
		2	0.531	0.308	0.003	0.220**
	Decoding	1	0.284	0.161	0.112**	0.009
		2	0.292	0.159	0.023*	0.095**
	Comprehension	1	0.181	0.079	0.002	0.063**
		2	0.321	0.168	0.007	0.084**

*P<.05 is marked

**p<.001 is marked.

Overall R2 includes contribution of oral vocabulary (PPVT).