Cognition 146 (2016) 453–469

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Adults with poor reading skills: How lexical knowledge interacts with scores on standardized reading comprehension tests

Gail McKoon *, Roger Ratcliff

The Ohio State University, United States

ARTICLE INFO

Article history:
Received 4 December 2014
Revised 8 October 2015
Accepted 12 October 2015
Available online 9 November 2015

Keywords:
Struggling adult readers
Diffusion modeling
Lexical decision
Reading scores

ABSTRACT

Millions of adults in the United States lack the necessary literacy skills for most living wage jobs. For students from adult learning classes, we used a lexical decision task to measure their knowledge of words and we used a decision-making model (Ratcliff's, 1978, diffusion model) to abstract the mechanisms underlying their performance from their RTs and accuracy. We also collected scores for each participant on standardized IQ tests and standardized reading tests used commonly in the education literature. We found significant correlations between the model's estimates of the strengths with which words are represented in memory and scores for some of the standardized tests but not others. The findings point to the feasibility and utility of combining a test of word knowledge, lexical decision, that is well-established in psycholinguistic research, a decision-making model that supplies information about underlying mechanisms, and standardized tests. The goal for future research is to use this combination of approaches to understand better how basic processes relate to standardized tests with the eventual aim of understanding what these tests are measuring and what the specific difficulties are for individual, low-literacy adults.

1. Introduction

The number of adults in the United States who have only the lowest of literacy skills is staggeringly high (The National Center for Education Statistics; Baer, Kutner, & Sabatini, 2009; Greenberg, 2008; Kutner, Greenberg, & Baer, 2006; Miller, McCardle, & Hernandez, 2010). The International Adult Literacy Survey Institute (2011) found that about 23% of adults in the United States read prose at the lowest level scored, indicating difficulty with comprehending even the most basic textual information; the National Assessment of Adult Literacy (Kutner et al., 2006) found that 43% lack the necessary literacy skills for most living wage jobs; and the Organization for Economic Co-operation and Development (OECD, 2013) found that one in six adults, about 36 million (two-thirds of them born in the United States) have low literacy skills (the comparable figure for Japan, for example, is one in 20). As Nicholas Kristof of the New York Times put it recently (October 26, 2014), these data "should be a shock to Americans." The Institute of Education Sciences in the United States Department of Education has made research to understand the skills these adults lack and how to teach those skills a high priority for funding (e.g., Calhoon, Scarborough, & Miller, 2013; Miller et al., 2010). The study we report here was designed to examine the viability of one new approach to the reading comprehension problems of this population.

We used a simple lexical decision task that is often used to study word comprehension, a skill that must figure largely in reading comprehension. In the lexical decision task, participants are given strings of letters and asked to decide as quickly and accurately as possible for each string whether it is or is not a word. For college undergraduates, accuracy on this task is typically above 90% and response times (RTs) average around 700 ms. The participants in our study were students in Adult Basic Learner classes with reading comprehension levels from the fourth through seventh grades. To their data, we applied a widely-accepted model for decision-making that decomposes RTs and accuracy into the cognitive mechanisms that underlie performance, namely, Ratcliff's diffusion model (Ratcliff, 1978; Ratcliff & McKoon, 2008).

One question was which mechanisms are deficient for low-literacy readers. Another was whether the model-based analyses we conduct can give insights into performance on a standardized language placement test for low-literacy adults, the TABE (Test of Adult Basic Education). A more general aim was to provide a proof-of-concept that diffusion model analyses are capable of informing practical education issues.

* Corresponding author at: Department of Psychology, The Ohio State University, Columbus, OH 43210, United States.
E-mail address: mckoon.1@osu.edu (G. McKoon).

http://dx.doi.org/10.1016/j.cognition.2015.10.009
0010-0277/

In the diffusion model, the information encoded from a stimulus is accumulated over time from a starting point to a criterion (a boundary), at which time a response is executed. For lexical decision, information accumulates toward a word boundary for "word" responses and toward a nonword boundary for "nonword" responses (Ratcliff, Gomez, & McKoon, 2004). Central to the model is that the accumulation of information is noisy – at any instant of time, the process may move toward one of the boundaries or it may move toward the other, but on average, a process will move to the word boundary for strings of letters that are words and to the nonword boundary for strings that are not. However, the noise is large enough that the incorrect boundary can sometimes be reached, resulting in an error, and that responses for the same item can reach a boundary at different times.

The model splits the decision process into three main components of processing. One is the settings of the boundaries, that is, how far they are from the starting point; this is assumed to be under the control of the individual making the decision (instructions to a participant or payoffs for one response over the other lead to adjustments in boundary settings, Ratcliff, Thapar, & McKoon, 2001, 2003; Ratcliff, Gomez et al., 2004). Another component is the quality of the information encoded from a stimulus, which determines the rate at which information is accumulated and is called "drift rate." For lexical decision, the quality of encoded information is determined mostly by the strength with which a word is represented in lexical memory (e.g., the representation of common words is stronger than the representation of rare words and so the rate of accumulation would be faster for common words). The third component is made up of processes outside the decision process itself: the time to execute a response and the time to encode a stimulus and transform it into a representation to drive the accumulation process. These processes are combined into one parameter of the model called "nondecision" time. In the model, the same decision process – from starting point to a boundary – determines the rate at which information is accumulated and which boundary will be reached.

For the purposes of this article, drift rates are the most interesting component because they measure the quality of the information about a word that an individual knows. This offers a new level of analysis for low-literacy research in three ways. First, often studies investigate correlations between individual-difference variables such as scores on tests of short-term memory, phonemic decoding, vocabulary, and standardized tests. Some of these tests have aspects of accuracy, the number of responses correct, and time, the amount of time given to produce responses, but none of these measure directly an individual's knowledge of words in the way the lexical decision task with a diffusion model analysis does. Second, the model can be applied to commonly used tests like those just mentioned. For example, short-term memory could be tested in a paradigm that asks individuals to decide whether or not a word was present in a just-presented list of words or vocabulary could be tested in a paradigm that asks individuals to decide which of two choices is the better match to a word's meaning. Paradigms like these could break performance into the components of processing defined by the diffusion model. Third, the model has been used to assess what readers know about the texts they read, for example, what the referent of a pronoun is, what the relations among elements of a text are, what the appropriate information to be inferred from a text is, and what the relations between information in a text and memory are (see McKoon & Ratcliff, 2015).

In performing lexical decision, and many other tasks, individuals can trade accuracy for speed or speed for accuracy. They can make their responses faster by setting their boundaries nearer the starting point, thus increasing the probability that the accumulated information will reach the wrong boundary. They can make their responses more accurate by setting their boundaries farther apart, thus making their responses slower.

In studies with low-literacy adults in the education literature, the speed/accuracy tradeoffs that individuals adopt and how these tradeoffs relate to underlying components of processing have not been explicitly considered. Understanding these tradeoffs is essential: An individual may respond with low accuracy to test items because the quality of the information encoded from the items is poor or because the quality of the information is good, but the boundaries are set close together. An individual may respond slowly to test items because the quality of the encoded information is poor or because it is good but the boundaries are set far apart. Another way to say this is that individuals with the same speed may have differences in accuracy and therefore differences in underlying mechanisms, and individuals with the same accuracy may have differences in speed and therefore differences in underlying mechanisms. It is these considerations that require speed and accuracy to be explained in concert and it is these considerations that require a model like the diffusion model to separate an individual's boundary settings from the quality of the information he or she encodes from a stimulus.

The importance of this separation is illustrated by applications of the diffusion model in aging research. Ratcliff et al. (2001, 2003), Ratcliff, Gomez et al. (2004), Ratcliff, Thapar, and McKoon (2007, 2010, 2011) have found that the usual aging effect – slower responses for older adults – often comes about not because the quality of the information they obtain from stimuli is less (i.e., not because their drift rates are lower) but instead because their nondecision component is slower and because they set more conservative boundaries, requiring more information to be accumulated before executing a response (e.g., Starns & Ratcliff, 2010). Thus, the frequently stated conclusion that older adults' cognitive processes are, overall, worse than young adults' because all cognitive processes are slowed is incorrect. In lexical decision, for example, older adults' drift rates have been as good or better than young adults' (Ratcliff, Thapar, Gomez, & McKoon, 2004).

In the sections below, we discuss research in education with low-literacy adults and research in cognitive psychology on word comprehension, then present the diffusion model in detail, and then describe the study we conducted.

2. Examples of multivariate research in the education literature

Many individual-difference studies in the education literature with low-literacy adults have used a psychometric approach to explore basic constructs that might contribute to the ability to understand written words. To illustrate this approach, we use three, quite recent, examples, studies by MacArthur, Konold, Glutting, and Alamprese (2010), Mellard, Fall, and Woods (2010), and Mellard, Woods, Desa, and Vuyk (2013; see also Nanda, Greenberg, & Morris, 2010, Tighe & Schatschneider, 2014). We discuss these in some detail to show the exploratory nature of the studies and to compare them to the diffusion model. The examples illustrate how psychometric approaches can differ from the diffusion model we use in this article. Those approaches look for broad, general constructs and how they are related to each other whereas the diffusion model provides analyses of the basic cognitive mechanisms that underlie comprehension skills. In other words, for constructs like those used in the three example studies, it would be possible, in principal, to use diffusion-model-like analyses to attempt to understand the mechanisms that determine performance.

In the 2010 study by Mellard et al., which had 174 participants in adult literacy classes, it was hypothesized that there are seven constructs relevant to reading comprehension: rapid automatic

naming, phonemic decoding, auditory working memory, word reading, reading fluency, vocabulary, and comprehension of spoken language. Information was assumed to flow among them from early processes to later ones to reading comprehension. To measure comprehension, participants were asked to read short passages and fill in a word that was left blank (a close procedure). They hypothesized that over 20 of the paths among the various tasks used to measure the constructs might be significant contributors to reading comprehension.

The model was tested by path analysis (a method related to multiple regression), which represents hypothesized relationships ("pathways") among variables. The method determines which coefficients (i.e., which pathways) among variables are significant. Mellard et al. found only 13 such pathways. For example, although the path from auditory working memory to spoken language comprehension was significant, other expected paths were not – neither rapidly naming letters nor phonemic encoding nor fluency contributed significantly to comprehension of spoken language.

MacArthur et al. (2010) divided reading-related skills into five clusters, each with tasks to measure it: decoding (three tasks), word recognition (three tasks), spelling (two tasks), fluency (two tasks), and comprehension (one task). The participants in this study were 486 students from Adult Basic Learner classes with reading comprehension levels from the fourth through seventh grades. The scores from the 11 tasks were entered into confirmatory factor analysis in order to group the tasks according to similarity in individual differences. Analyses showed that a five-cluster model explained the correlational structure of the tasks better than two- or three-cluster models and so it was concluded that abilities corresponding to the five clusters should be assessed separately for studies of low-literacy individuals.

Mellard et al. (2013) used a third approach in describing the skills of 290 low-literacy adults who were Job Corps students. They hypothesized six constructs as determinants of reading comprehension: phonological processing, word reading, spelling, vocabulary, processing speed (fluency), and cognitive ability. For each of the constructs, there were three tasks. For example, processing speed was measured with a task for which the students were given a row of pictures and asked to quickly identify the two most conceptually similar.

Mellard et al. (2013) applied principal axis factoring to the data. This analysis method is different from the two above in that it does not impose a preconceived structure; rather it identifies the structure and relationships among variables. It produces clusters of variables such that each cluster is made up of the variables that are correlated with each other more than with any of the others. Then, with these clusters, Mellard et al. used multiple regression to see how well performance was predicted on two standardized reading tests.

The results showed only four clusters, not the hypothesized six. Mellard et al. (2013) labeled the four clusters as encode/decode (seven of the 18 tasks), vocabulary (five of the tasks), processing speed (three of the tasks), and working memory (four of the tasks). The tasks were also organized differently than Mellard et al. had expected. For the encode/decode cluster, reading single words and spelling were grouped together; for the vocabulary cluster, the three expected tasks plus two phonological coding tasks were grouped together; for the processing speed cluster, the two processing speed tasks plus word reading were grouped together; and for the working memory cluster, the expected three working memory tasks and one of the others were grouped together. In addition, some of the results were odd, for example, the task of blending spoken sounds together to make a word and the task of completing spoken words with missing letters were not grouped with the tasks in the encoding/decoding cluster.

These studies give valuable information about how the various tasks and constructs may be related through individual differences and this is information that can be used to develop hypotheses to guide further research. The studies provide topics for further research such as how and why tasks are related to each other and how and why they are related to hypothesized constructs. However, currently, interpretations of results like those just reviewed cannot be seen as being completely settled. If low-literacy individuals do well on some tasks, it is tempting to conclude that the construct at which the tasks are aimed is not problematic for them. But, as noted, constructs can be turned into experimental tasks in many ways and successful performance on one or two of them does not necessarily mean successful performance on others. Similarly, if low-literacy adults fail to show expected relations among constructs, it could be that the tasks used to test the constructs were the wrong ones to show the expected relationship, that there was no relationship, that there was not sufficient power in the tasks to detect relationships, or that the tests have not been validated for low-literacy adults.

3. Cognitive psychology approaches

There is a huge amount of research on word comprehension in experimental cognitive psychology, especially with adults and children who have reading disabilities such as dyslexia. This research differs from the diffusion model approach in two important ways. First, much of this research has used only RT as the dependent variable. However, as discussed above, any explanation of behavior should explain speed and accuracy simultaneously; they must come from the same underlying mechanisms. A second difference is that, while most models in this area can explain the effects of independent variables qualitatively, they have not been shown to fit experimental data and so explain them quantitatively. Furthermore, because they are not fit to data quantitatively, they are not able to estimate parameters of processing for individual participants and so cannot be used to examine the effects of and relationships to individual differences such as IQ.

Many studies of word comprehension difficulties have used the lexical decision task (e.g., Grainger & Jacobs, 1996; Seidenberg & McClelland, 1989) to examine a host of independent variables. To name just a few, responses are easier for words that occur frequently in English than those that occur less frequently; responses to a word are more difficult if there are many words similar to it (e.g., back is more difficult than most words because there are many words similar to it: buck, tuck, luck, lack...); responses to words learned early in life are easier than those learned later; and words with clusters of letters in them that occur frequently are easier than those with clusters that occur less frequently. There are also semantic priming effects such that a response to a word is speeded if it is immediately preceded in a test list of strings of letters by a semantically related word (e.g., nurse preceded by doctor). Importantly, the fact that these variables affect lexical decision performance means the lexical decision task taps the representations and meanings of words in lexical memory.

There are several models that have attempted to explain the effects of all of these variables on the strength with which a word is encoded in lexical memory (Grainger & Jacobs, 1996; McClelland & Rumelhart, 1981; Perry, Ziegler, & Zorzi, 2007; Seidenberg & McClelland, 1989). For example, all other things being equal, the strength of high-frequency words should be greater than the strength of low-frequency words. In the models, the various processes that determine the strength of a word operate with both top-down and bottom-up processes, and all of the disparate processes come together to support word recognition in just a few hundred ms. Higher level processes support lower level processes

and lower level processes support higher level processes, and processes that operate in parallel are integrated with processes that operate serially.

Models like these can produce values of the strengths with which words are represented in lexical memory but they have not linked this strength with a decision model that explains accuracy and RTs in the lexical decision task. Ideally, a full model would take output from a word comprehension model and use that to drive a decision model, but this has rarely been attempted (although see Dufau, Grainger, & Ziegler, 2012; Norris, 2006; Ratcliff, Gomez et al., 2004).

Often, when word comprehension models have been used to explore the locus of word comprehension disabilities such as dyslexia (e.g., Harm, McCandliss, & Seidenberg, 2003; McLeod, Shallice, & Plaut, 2000; Patterson, Seidenberg, & McClelland, 1989; Ziegler et al., 2008), the model for normal comprehension is damaged by, for example, reducing the numbers of units in the hidden layers of a connectionist model, and then the behavior of the damaged model is compared to the behavior of patients. However, as noted above, these models have not been explicitly fit to data from individual participants and so it has not been demonstrated that these models can measure effects at the level of individual participants in experiments.

4. The two-choice diffusion model

The model is intended to explain all of the data from two-choice tasks like lexical decision for which responses are made in under a second or two: accuracy, the mean RTs for correct responses, the mean RTs for incorrect responses, and the full distributions of RTs for correct and incorrect responses, and it is intended to explain speed and accuracy simultaneously with the same decision process. It has been successful in doing all of this in many studies (see examples in Ratcliff and McKoon (2008)).

The model is illustrated in Fig. 1. The noisy information encoded from a stimulus, in this case, from a string of letters, accumulates from the starting point, "z," toward the decision boundaries. In the top panel of Fig. 1, the arrow shows the mean rate of approach, that is, drift rate, to the word boundary for a letter string that is a word. With this drift rate, the arrow is heading fairly steeply toward the word boundary and so most responses would be fairly fast "word" responses. The drift rates for easier words (e.g., common words) are higher than those for more difficult words (e.g., rare words). Letter strings that are nonwords have negative values of drift rate. Their absolute values are larger for letter strings that are quite different from real words than for strings that are similar to real words.

The total processing time for a decision is the sum of the time taken by the decision process to reach a boundary and the time taken by the nondecision component (middle panel of Fig. 1). The bottom panel of Fig. 1 illustrates the role the model plays in relating performance to underlying mechanisms. Accuracy and the distributions of RTs for correct and error responses for each condition of an experiment map through the model to drift rates, boundaries, and nondecision times.

The three paths in the top panel of Fig. 1 have the same mean drift rate but, because of the noise in the accumulation process, they lead to different outcomes. One leads to a fast correct decision, one to a slow correct decision, and one to an error. This noise is responsible for the shapes of RT distributions, as shown in the figure. Most responses are reasonably quick responses, but there are slower ones that spread out the right-hand tails of distributions. The third path in the figure illustrates that even when drift rate is strongly positive, the accumulation of information can reach the negative boundary.

Each component of the model – drift rate, the settings of the decision boundaries, and the nondecision component – is assumed to vary across the trials of an experiment. The idea is that participants in an experiment cannot hold the values exactly constant from one trial to the next. This is called "between-trial" variability, contrasting it with the "within-trial" noise in the accumulation process (i.e., the noise illustrated in Fig. 1). The assumption of across-trial variability allows the model to account for the relative speeds of correct and error responses (see Ratcliff & McKoon, 2008).

Speed/accuracy tradeoffs are handled mainly by the boundary settings, not drift rates or the nondecision component. The separation of boundary settings and nondecision times from drift rates is fundamental to the model. To illustrate, suppose that for skilled readers, lexical decision RTs average 700 ms and accuracy averages 95% and for poor readers, RTs average 900 ms and accuracy averages 98%. The question is whether drift rates could be the same for the two groups even though the poor readers are considerably slower but 3% more accurate. In fact they can. These data were simulated from the model with drift rate set to 0.45 (fairly strong toward the word boundary) for both groups. In the simulated data, the poor readers differed in RTs and accuracy only because they set their boundaries farther apart.

As discussed in the introduction, the model provides an understanding of why accuracy and RT are not correlated across subjects. This results from the separation of drift rates, boundaries settings, and nondecision times, which are (largely) independent components of processing. Accuracy is mostly determined by drift rates and RTs mostly by boundary settings and to a lesser degree by nondecision time (and sometimes slightly by drift rates). The separation explains why results from studies that measure only RT can lead to conclusions that conflict with those from studies that measure only accuracy (e.g., the studies with older adults cited above). The former are determined more by boundaries and nondecision times and the latter more by drift rates.

A key feature of the model is that, when it is fit to data, the variability in the estimate of a component is substantially less than the variability among individuals. This means that it can measure components of processing at the level of an individual, which has led to interpretable differences among individuals in a number of applications (e.g., Ratcliff, Schmiedek, & McKoon, 2008; Ratcliff, Thapar, & McKoon, 2006, 2010, 2011; Ratcliff, Thompson, & McKoon, 2015; Schmiedek, Oberauer, Wilhelm, Suβ, & Wittmann, 2007; Spaniol, Madden, & Voss, 2006; Wagenmakers, Van Der Maas, & Grassman, 2007).

Fitting the model to data is accomplished by a minimizing routine that iteratively adjusts parameter values (drift rate, boundaries, the nondecision component, and the variability across trials in each of them) until the values that best predict the data are obtained (see Ratcliff & Tuerlinckx, 2002, for details). For each condition in an experiment, quantile RTs are generated and these, through the model, generate the proportions of predicted responses between the quantiles. These proportions multiplied by the number of observations are used to produce a chi-square value and the model parameters are adjusted to make the chi-square value (summed over conditions and correct and error responses) a minimum. Recently, three software packages have been developed and they are in wide use (Vandekerckhove & Tuerlinckx, 2008; Voss & Voss, 2007; Wiecki, Sofer, & Frank, 2013; see Ratcliff & Childers, 2015, for an evaluation). However, we used Ratcliff's routines for the experiment here because we can adapt them more quickly and easily than the packages.

Essential to a model is that it be identifiable and falsifiable. While it is relatively easy for the diffusion model to fit mean RTs and accuracy, and it can do so with a range of different parameter values (i.e., it would not be identifiable), it is severely constrained

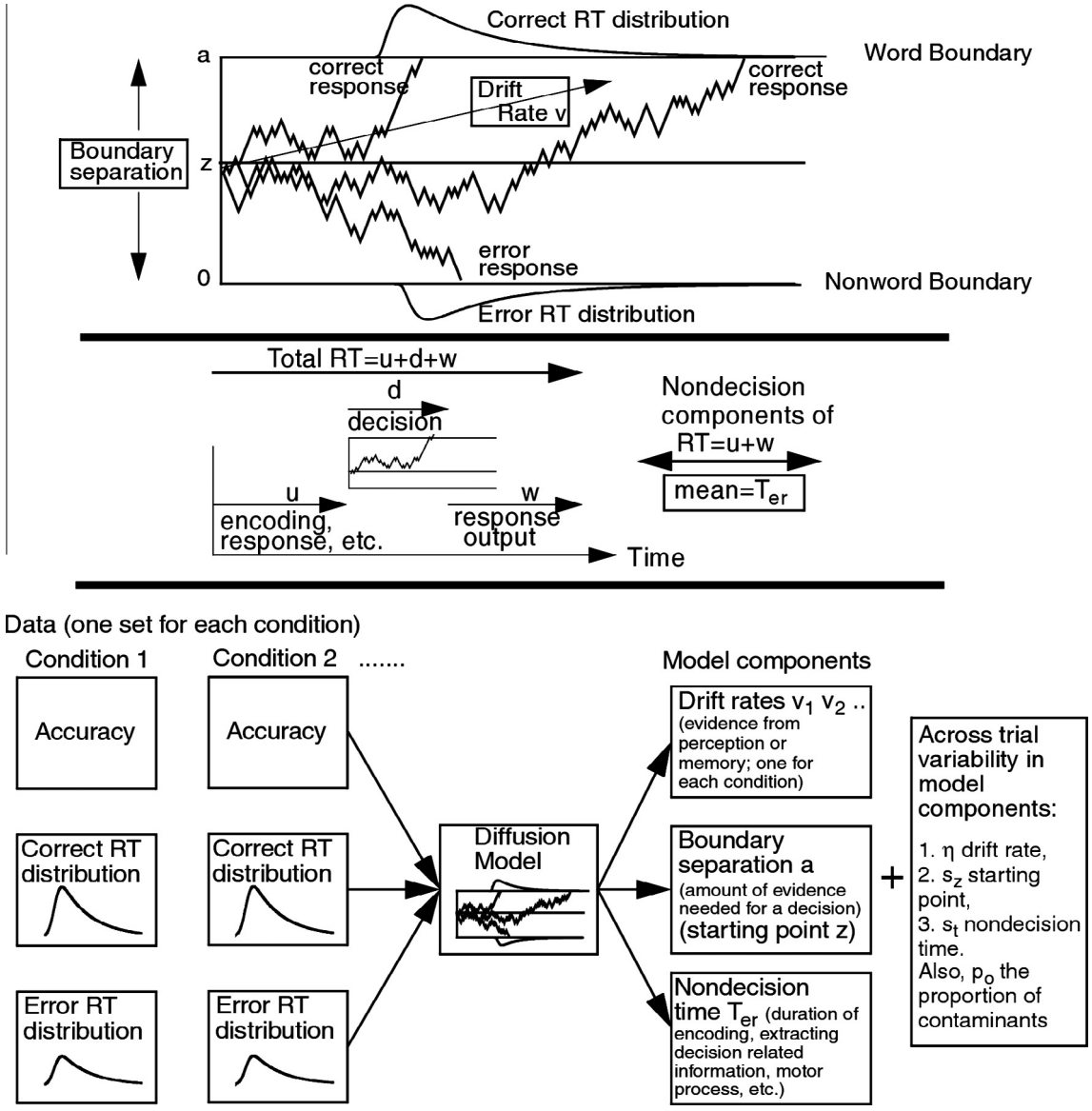

Fig. 1. An illustration of the diffusion process. The top panel shows three simulated paths with drift rate v, starting point z, and boundary separation a. Drift rate is normally distributed with SD η and starting point is uniformly distributed with range s_z. Nondecision time is composed of encoding processes, processes that turn the stimulus representation into a decision-related representation, and response output processes. Nondecision time has mean T_{er} and a uniform distribution with range s_t. The bottom panel illustrates the mapping between accuracy and RT distributions to diffusion model parameters.

by the shapes of RT distributions. Ratcliff (2002) made up several sets of fake but quite plausible data and showed that the diffusion model failed (dramatically) to fit them. Jones and Dzhafarov (2014) have recently claimed that the model is not falsifiable, but their claim is valid only for models in which there is no or almost no within-trial variability (see Smith, Ratcliff, & McKoon, 2014 for detailed discussion). Also, in most comparisons made so far, e.g. Ratcliff, Thapar, Smith, and McKoon (2005), we have found similar interpretations of data from competing models (e.g., Usher & McClelland, 2001; see also Donkin, Brown, Heathcote, & Wagenmakers, 2011).

It is important to point out two distinctions between the empirical methods used for diffusion model analyses and those used for many standardized tests like the TABE or the TOWRE tests. The diffusion model applies to decisions that are made quickly, under a second or two. They are "one-shot" decisions in that they involve only one decision process, not multiple attempts to encode stimuli and consider which response to make. With RTs short, the model assesses "automatic" processes, not "strategic" ones. Automatic processes occur quickly, passively, without conscious effort and strategic processes take more time and involve some effort. Automatic processes determine much of the information understood during reading (e.g., the contextually appropriate meanings of words, the referents of pronouns) and they are mainly responsible for "moment-to-moment" comprehension. This is reflected in, for example, college students' average reading rate, which is about 250 ms per word. Lexical decision has been shown to rely primarily on automatic processes (Neely, 1977; Posner & Snyder, 1975a, 1975b; Ratcliff & McKoon, 1981). Tests like the TABE and TOWRE have a more strategic nature.

The second distinction is that speed of processing is handled quite differently in the diffusion model than in tasks for which the measure is the number of test items that an individual completes in a fixed amount of time. Performance on the latter involves

speed, but speed cannot be separated from the quality of the information that guides decisions. Also, it may not measure the same aspect of speed that rapid RT tasks measure.

5. Experiment

The population of adults with poor reading skills is extremely heterogeneous (e.g., Greenberg, 2008). In addition to wide variations in age, socio-economic status, ethnicity, and employment history, there are any number of reasons that individuals may not have learned to read well, including, for example, financial difficulties, lack of motivation, peer pressure, deficits such as dyslexia, or non-native home languages. As a result, differences among individuals in various components of reading can be quite large even when they read at approximately the same level as measured by some standardized test. This is one of the motivations for analyzing reading performance at the level of individuals.

The population of low-literacy adults also presents a challenge for the diffusion model. The model has been developed with a typical undergraduate population; undergraduates are easily available, easy to test, and usually have good reading skills. The diffusion model has also been successful with a number of other populations, including children (Ratcliff, Love, Thompson, & Opfer, 2012), sleep-deprived individuals (Ratcliff & Van Dongen, 2009), aphasic individuals, (Ratcliff, Perea, Colangelo, & Buchanan, 2004), hypoglycemic individuals (Geddes et al., 2010), children with ADHD (Mulder et al., 2010), children with dyslexia (Zeguers et al., 2011), individuals with anxiety, and individuals with depression (White, Ratcliff, Vasey, & McKoon, 2009; White, Ratcliff, Vasey, & McKoon, 2010a, 2010b). However, individuals in these populations either have no significant deficits in reading skills or have major diagnosed problems, so it is an open question whether the model can be successful with adults with non-specific low reading skills.

In the experiment, there were 124 students from Adult Basic Literacy Education (ABLE) classes in the Columbus, Ohio, area, all native speakers of English, ranging in age from 18 to 78. These classes are free and anyone can attend. We collected six individual difference variables: age and performance on the TABE, WAIS vocabulary IQ, WAIS matrix reasoning IQ, the TOWRE test for words, and the TOWRE test for nonwords.

The TABE is used to assign reading levels for students in ABLE classes. It is a widely used test of adults' reading ability (Beder, 1999; Ehringhaus, 1991). It consists of a series of texts, each with several paragraphs, on topics such as household cleaners, cell phone purchasing plans, and "The Power of Color," with several multiple choice questions for each.

The WAIS tests are untimed tests. The vocabulary test asks participants to give definitions of 33 words. The words increase in difficulty from beginning to end and testing is stopped when a participant cannot give a definition for six words in a row. The score is the number of points earned, with each word worth a maximum of two points (one point is awarded for a partially-correct definition). For the WAIS matrix reasoning test, each of 26 items is made up of a logic puzzle with a missing piece and participants are asked to choose from five choices which would be the one that completes the puzzle. The items increase in difficulty from beginning to end and testing is stopped when a participant is incorrect for four items in a row or four out of five items in a row. The score is the number of correct responses.

The TOWRE test for words contains 108 words and the TOWRE test for nonwords contains 66 nonwords (Torgesen & Wagner, 1999). For both, the items increase in difficulty from the beginning of the list to the end. The score is the number of items pronounced correctly in 45 s.

We note that scores on the TABE have extra variability because different individuals receive different versions of the test. Individuals first take a "locator test," which determines whether they will be tested on the easy, medium, difficult, or advanced version of the TABE, and then their score on the TABE is scaled to give their reading grade level. The result is that someone at a grade level of 4.6, for example, could have achieved that from any of the four versions. This is unlike the TOWRE and WAIS tests in which a single instrument is used.

The TOWRE tests and the TABE test were chosen because they are frequently used in educational research with low-literacy adults. Age and the two IQ measures were chosen because they significantly affect performance on many cognitive tasks. For the standardized tests, the same items are used for every individual who takes them, with the aim of reducing variability among individuals. For lexical decision (and many other tasks in cognitive psychology), words are drawn from a large pool with each participant getting a different random sample from the pool. The aim is to generalize findings across as many words as possible.

For this experiment, we did not generate predictions about which individual-difference measures and which components of the model would be correlated; in this sense, the experiment was largely exploratory. However, the model has a significant bar to pass: interpretations of data cannot proceed until it has successfully accounted for the data (i.e., accuracy and the distributions of RTs for every condition in the experiment). In this way, it contrasts with the individual-differences research from education that we described above where scores on various tests are entered directly into analyses (e.g., multiple regression, path analysis).

There were also 56 undergraduate students in the experiment. The individual difference measures were not collected for them; they were included in the experiment only to provide a benchmark against which the ABLE students' RTs, accuracy, and components of the diffusion model could be compared.

6. Method

6.1. Materials

Overall, 4211 words were tested for lexical decision. Each participant was tested on only a subset of the words, with the subsets ranging from 660 words to 1040 words. For each subset, the number of nonwords tested was equal to the number of words and the nonwords were matched to the words in terms of number of letters. The words and nonwords were presented in random order.

We divided the words into two sets that were chosen on the basis of the results of the experiment, 3641 for which accuracy was above 80% for the ABLE students and 570 for which accuracy for them was 80% or below (mean 68% for ABLE students, 85% for undergraduate students). We call the former "good words" and the latter "bad words" (we use the term "bad" to be evocative and easy to remember). The experiment also served the purpose of determining whether a word is well known or not to the ABLE students and so could be used or not in later text processing experiments. For the good words, the frequency of occurrence in English (Kucera-Francis) averaged 49.2 and for the bad words, it averaged 14.3. 189 bad words were repeated in three of the subsets (i.e., with three groups of participants – every participant was tested on this same set of 189 words) and it was these that we used for the analyses reported below. As the results below show, the separation of bad words from good words gave more information about ABLE students' performance than if all the words were analyzed together.

Nonwords and bad words were two conditions of the experiment. In order to provide an extra condition (to give more con-

straints on the diffusion model), we divided the good words into two conditions, 2069 medium-frequency words and 1571 low-frequency words. The mean frequency (Kucera-Francis) for the medium-frequency ones was 81.2 (with SD 89.8) and the mean for the low-frequency ones was 6.8 (with SD 5.8).

6.2. Procedure

For the lexical decision task, letter strings were displayed on the screen of a PC. Participants were asked to respond to each one as quickly and accurately as possible, indicating a "word" response by pressing one key on the PC keyboard and "nonword" with another key. Each letter string was displayed until a participant made a response. If the response was correct, there was a blank screen for 50 ms, and then the next string of letters was displayed. If the response was incorrect, then the word ERROR was displayed for 900 ms, then a blank screen for 50 ms, and then the next string. After between 44 and 60 trials (depending on the number of words being tested), participants could take a break and then press the space bar on the keyboard to continue.

6.3. Participants

For the 124 students from ABLE classes, means and standard deviations of the individual difference measures are shown in Table 1. The 56 undergraduates were students at Ohio State University who participated in the experiment for credit in an undergraduate psychology course. None of the individual difference measures were collected from them.

7. Results

7.1. Data

For lexical decision, responses longer than 4000 ms and shorter than 400 ms were excluded from analyses (about 2% of the data). Mean RTs and accuracy values were calculated for all the participants for all the words. These values and their standard deviations are shown in Table 2.

7.2. The diffusion model fit the data well

The model accounted well for the data from the ABLE students and the undergraduates, just as it has in previous studies with lexical decision with other populations (Ratcliff, Gomez et al., 2004; Ratcliff, Thapar, Gomez et al., 2004; Ratcliff et al., 2010). Plots of experimental versus predicted values of accuracy and the 0.1, 0.5, and 0.9 quantiles of RTs for correct responses are shown in Fig. 2. For ABLE participants, each plot contains 124 participants by four experimental conditions (480 points per plot) and for undergraduate participants, each plot contains 56 participants by four experimental conditions (224 points per plot). There are some misses but the proportion of them that are serious is small. The mean chi-square values, shown in Table 3, are near twice the critical value (47.4 with 33 degrees of freedom) but the chi-square is a conservative test and the number of observations per participant is large (1200–2000), which means that even small differences between theory and data lead to large values for chi-square (see Ratcliff, Gomez et al., 2004, for further discussion). Table 3 shows the means of the values of the parameters that produced the best fits to the data.

7.3. ABLE students compared to undergraduates

A central question for this study was whether ABLE students' performance was worse than undergraduates' (longer RTs and lower accuracy) because their knowledge of words was worse, as would be expected. As described above, the diffusion model can answer this question directly because it separates drift rates from boundary settings and nondecision times. In fact, the drift rates were significantly different between the ABLE students and the undergraduates, $F(1,178) = 67.1$, $p < .05$. However, not all of the difference in performance was due to drift rates; the ABLE students set their boundaries more conservatively (i.e., the distance between their boundaries was larger), $t(161) = 9.1$, $p < .05$, and their nondecision times were longer, $t(177.8) = 7.7$, $p < .05$.

We can take the question a step further by examining how much of the RT and accuracy differences were due to differences in boundaries and how much were due to differences in drift rates. To look at boundaries, we replaced the boundary separation (i.e., the value of a) and starting point (i.e., z) for the undergraduates with the boundary separation and starting point for the ABLE students. When we did this, accuracy increased by 0.8–3.5% depending on condition and mean RT increased by 81–180 ms. When we replaced the drift rates for the undergraduates with the drift rates for the ABLE students, accuracy decreased by 2–16% depending on condition and mean RT increased by 45–58 ms. This analysis shows that the reduction in accuracy from undergraduates to ABLE students was due to a reduction in drift rates and half the increase in RTs was due to boundaries, with the other half shared by nondecision time (a difference of 58 ms) and drift rates.

Comparing the two groups on the other components of the model, starting points were approximately equidistant from the boundaries for both groups and the three variability parameters were similar. We also modeled the proportions of "contaminant" responses, those that were likely due to processes other than those of interest, for example, distraction. To represent these trials, we assume that, on some proportion of trials (po), a uniform-distributed random delay between the minimum and maximum RT for the condition is added to the decision RT (see Ratcliff & Tuerlinckx, 2002). The estimates of the proportion of these contaminants were also similar between the ABLE students and the undergraduates.

7.4. Correlational analyses

In fitting the model to the data, the standard deviations in the parameter values were less than the standard deviations among

Table 1
Background measures ABLE students.

	Yrs edu	Age	Voc raw	Voc scal	Mat raw	Mat scal	IQ	TOWR word	TOWR nonwd	TOWR non strt	TABE	Grade level
Mean	9.80	41.7	24.1	6.09	11.30	8.11	83.2	75.07	36.30	32.97	3.80	6.89
min	5	18.6	7	2	4	3	63	34	10	7	1	1.9
max	12	77.8	54	13	24	16	123	106	62	62	6	12.9
SD	1.47	14.0	8.2	1.95	5.21	2.73	10.0	11.48	12.44	12.37	1.00	2.41

Yrs edu is the number of years of education completed; age is the subject's age; Voc raw, Voc scal, Mat raw, Mat scal are the raw and scaled scores on the WAIS-III (vocabulary and matrix reasoning subtests); IQ is estimated IQ; TOWR word and TOWR nonwd are the number of words and nonwords correct on TOWRE-2; TABE is the reading Educational Functioning Level from TABE; Grade level is the estimated reading grade level.

Table 2
Accuracy and mean RT.

	Low-frequency words		Medium-frequency words		Bad words		Nonwords	
	Accuracy	RT	Accuracy	RT	Accuracy	RT	Accuracy	RT
ABLE mean	0.934	955	0.957	873	0.686	1206	0.875	1059
Ugrad mean	0.962	698	0.979	659	0.849	798	0.946	749
ABLE SD	0.057	193	0.041	176	0.139	271	0.094	248
Ugrad SD	0.029	71	0.017	66	0.071	90	0.055	94

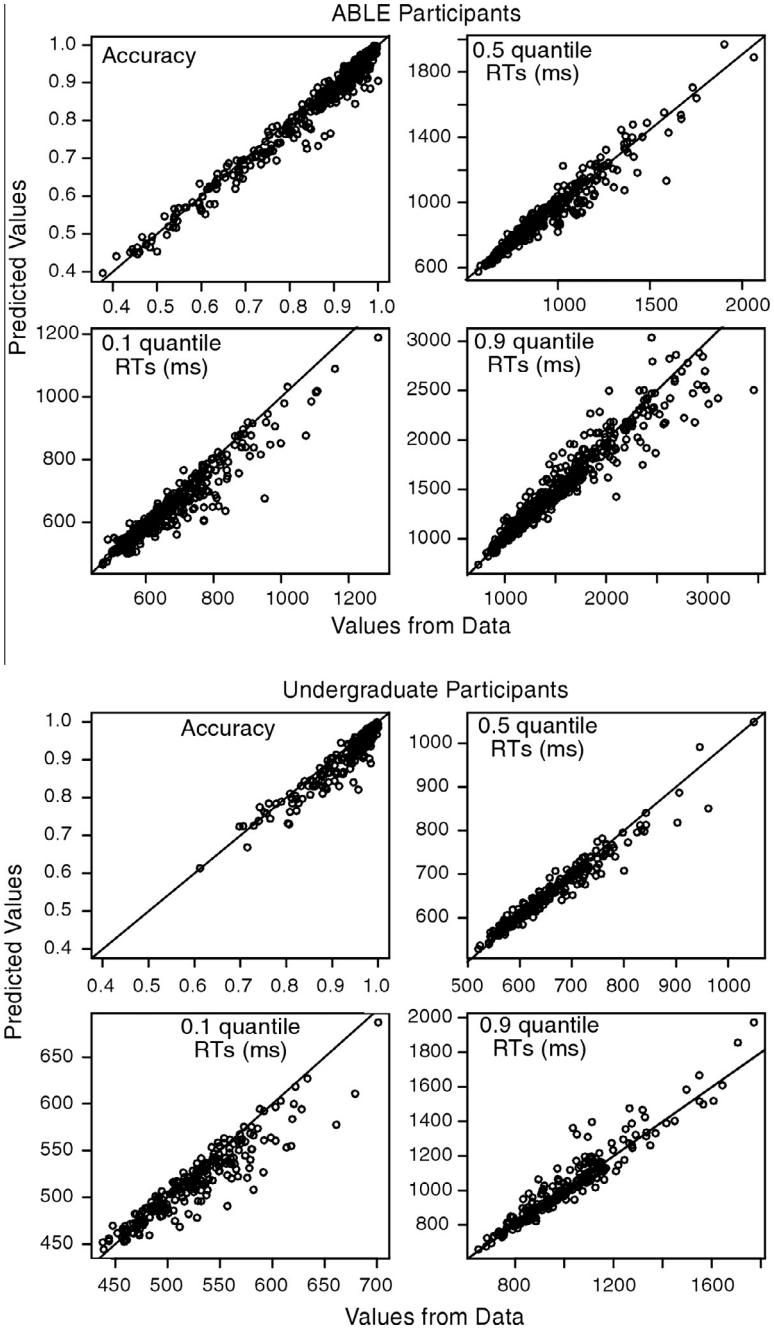

Fig. 2. Plots of accuracy and the .1, .5 (median), and .9 RT quantiles for data (*x*-axis) and predicted values from fits of the diffusion model (*y*-axis) for correct responses for ABLE students and undergraduates.

participants, which means that correlations among the values as well as correlations among the values and standardized test scores could be meaningfully calculated. The success of these analyses is especially noteworthy for two reasons. One is that the ABLE students had a limited range of abilities, all scoring between the fourth and seventh grade levels on the TABE and

Table 3
Diffusion model parameters.

	a	z	T_{er}	η	s_z	p_o	s_t	v_{low}	v_{med}	v_{bad}	v_{non}	χ^2
ABLE mean	0.193	0.095	0.512	0.103	0.075	0.015	0.164	0.232	0.295	0.065	-0.162	84.4
Ugrad mean	0.146	0.076	0.454	0.099	0.066	0.020	0.130	0.308	0.379	0.160	-0.269	79.0
ABLE SD	0.040	0.024	0.065	0.049	0.038	0.024	0.093	0.074	0.090	0.053	0.065	36.9
Ugrad SD	0.025	0.017	0.028	0.048	0.024	0.027	0.050	0.071	0.086	0.050	0.068	30.0

Note: a = boundary separation, z = starting point, T_{er} = nondecision component of response time, η = standard deviation in drift across trials, s_z = range of the distribution of starting point (z), p_o is the proportion of contaminants, s_t = range of the distribution of nondecision times, v = drift rates, and χ^2 is the chi-square goodness of fit statistic.

almost all scoring below 100 on the IQ tests. The other reason, for correlations that included the TABE, scores on it are merged from several different versions of it. As mentioned above, the power of the correlations might come, at least in part, from there being more heterogeneity in the ABLE population than small ranges of grade levels suggest (e.g., Greenberg, 2008; Keenan & Meenan, 2014).

A first important result for the ABLE students was that drift rates were not significantly correlated with boundary settings or nondecision times. That is, wherever an individual set his or her boundaries or whatever his or her nondecision time, the model provides a measure of the individual's knowledge of words. This is the signal contribution of the model.

Given that drift rates measure word knowledge, then it can be asked whether such knowledge is implicated in the other measures that we used as it should be. This was the case: drift rates correlated significantly with WAIS IQ vocabulary and TOWRE word and nonword scores. For TABE scores, it might be thought that understanding the complexity of a text would swamp the effect of knowledge of individual words, but TABE scores were significantly correlated with drift rates.

In the next paragraphs, we more fully describe five findings for the ABLE students.

(1) Fig. 3 shows histograms, scatter plots, and correlations for the finding of no significant correlations among drift rates, boundary settings, and nondecision times (with one weak exception, the correlation between drift rate and boundary separation). The figure also shows what would be expected from other diffusion model analyses: that boundary settings and nondecision times correlated significantly with RTs but not with accuracy and drift rates correlated significantly with accuracy and only weakly with RTs.

(2) Fig. 4 shows the significant correlations among all the word-related measures, drift rates, WAIS vocabulary IQ, the TOWRE scores, and TABE scores, and their histograms and scatter plots. For WAIS vocabulary IQ and matrix reasoning IQ, the scores are integer numbers and many of them are identical (they range between 2 and 16). If the values were plotted, 30 observations for a score of 10 would not appear different from 1 observation for a score of 10 (the 30 points would all lie on top of each other) and thus provide a false view of the correlations. Instead, plots should show the density of points and to do this, we added to each score a random normally distributed number with SD 0.2 of the score, i.e., we jittered the scores. This gives plots that show a dense region when there are many scores with the same value, as in Fig. 4. However, all of the correlations were performed on the unjittered data.

The correlations among WAIS vocabulary IQ, TOWRE scores, TABE scores, and drift rates were all larger for the bad words than the good words or the nonwords. It might be thought that the correlations for the good words or the nonwords were low because performance was at ceiling and so estimates of drift rates were uncertain. However, if ceiling effects were the problem then the correlation between nonwords and good words would be smaller than the correlations between either of them and bad words (because both nonwords and good words would have ceiling effect issues). Instead, drift rates for good words and nonwords correlated .74 with each other, a value that is larger than the correlation between bad words and good words (.54) and between bad words and nonwords (.40). Thus for ABLE students, scores lined up the same way across individuals for good words and nonwords (if someone had a high score on one, they had a high score on the other), but they lined up somewhat differently for bad words relative to either good words or nonwords.

(3) Age correlated significantly with boundary settings and nondecision times, but not drift rates, reflecting the standard finding that older subjects are slower than younger ones because of their conservative boundary settings and nondecision times, not because their information about words is of lesser quality (Ratcliff et al., 2001; Ratcliff et al., 2003; Ratcliff, Gomez et al., 2004; Ratcliff et al., 2010; Ratcliff et al., 2011). Fig. 5 shows histograms, scatter plots, and correlations among age, boundary settings, nondecision times, and drift rates, and correlations among them and TOWRE scores. There were no significant correlations among all of these variables and WAIS vocabulary or WAIS matrix reasoning scores. For the TOWRE tasks, the significant correlations between their scores and nondecision times are plausible: For lexical decision, nondecision time is the time taken up by processes outside the decision process and so it is reasonable that they have commonalities with the processes by which strings of letters are pronounced.

(4) Fig. 6 shows the histograms, scatter plots, and correlations among accuracy and RTs averaged over word and nonword conditions and WAIS vocabulary IQ, WAIS matrix reasoning IQ, TOWRE scores, and TABE scores. The correlations between accuracy and RTs and WAIS vocabulary IQ, TOWRE scores, and TABE scores were significant, although only modestly so for those that involved RTs. WAIS matrix reasoning IQ did not correlate significantly with accuracy or mean RTs. Fig. 6 also shows the non-significant correlation between speed and accuracy that we have discussed (.05, top left correlation and scatter plot). Also, in the second row of numbers in the top two lines, it shows the correlations for the bad words. These values are similar to the correlations for the mean RT and accuracy values averaged across conditions. Accuracy and mean RT for the medium- and low-frequency words and the nonwords gave correlations lower than the values in Fig. 6.

(5) WAIS matrix reasoning IQ was significantly (but only modestly) correlated with the TABE scores and WAIS vocabulary IQ, but not with anything else (Fig. 4). This suggests that performance on the lexical decision task, the TOWRE tests, and vocabulary IQ is at least somewhat

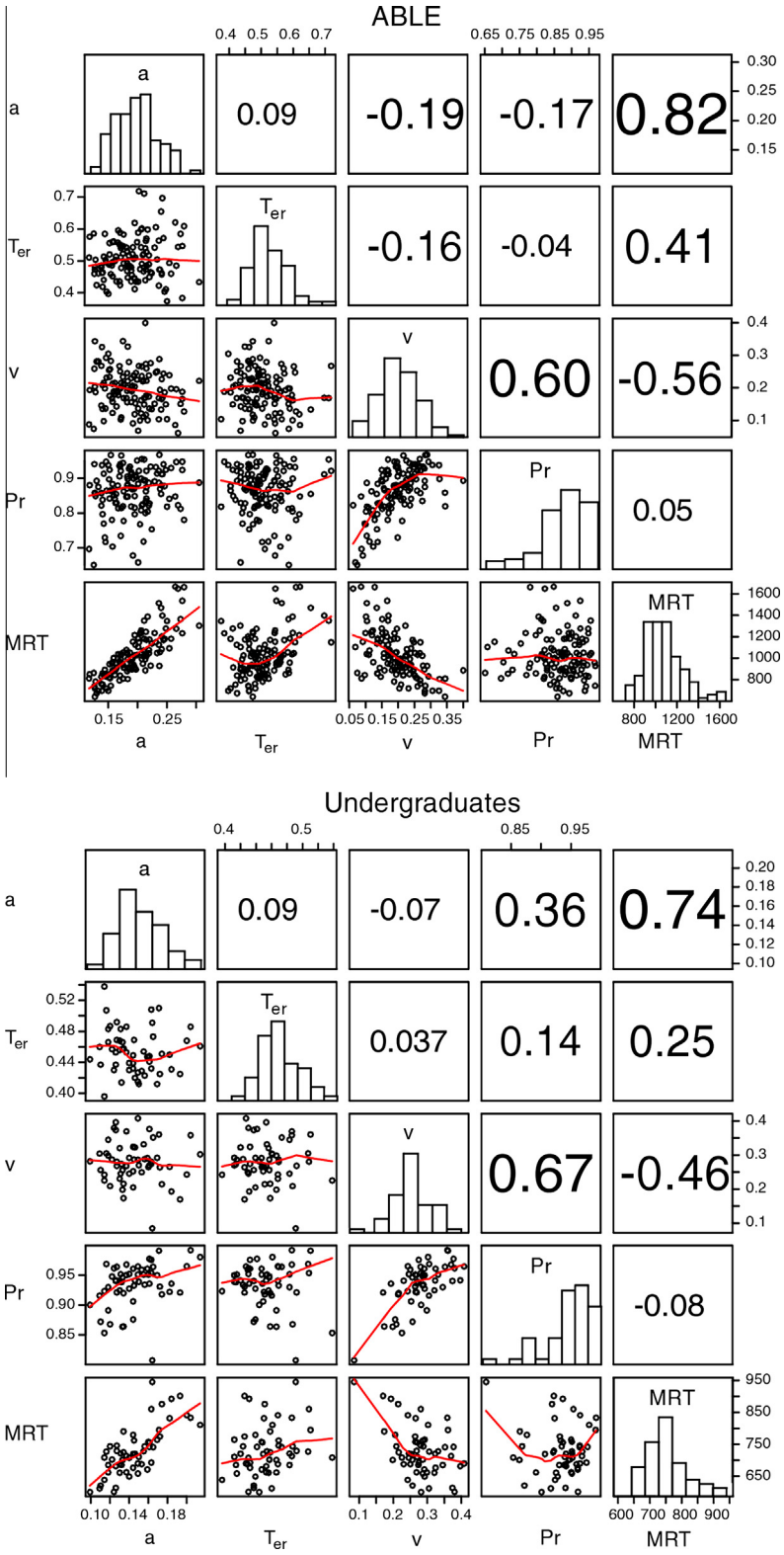

Fig. 3. Scatter plots, histograms, and correlations for boundary separation, nondecision time, mean drift rate across conditions, mean accuracy across conditions, and mean RT across conditions. The top panel is for ABLE students and the bottom panel for undergraduates.

separate from reasoning skills. The small but significant correlation between WAIS matrix reasoning IQ and TABE scores suggests that the TABE may involve reasoning skills as well as the word and reading skills represented by the other measures.

7.5. Correlations for the undergraduates

For accuracy, RTs, drift rates, boundary settings, and nondecision times, the patterns of the correlations (Fig. 3) were the same as for the ABLE students except that the undergraduates' boundary

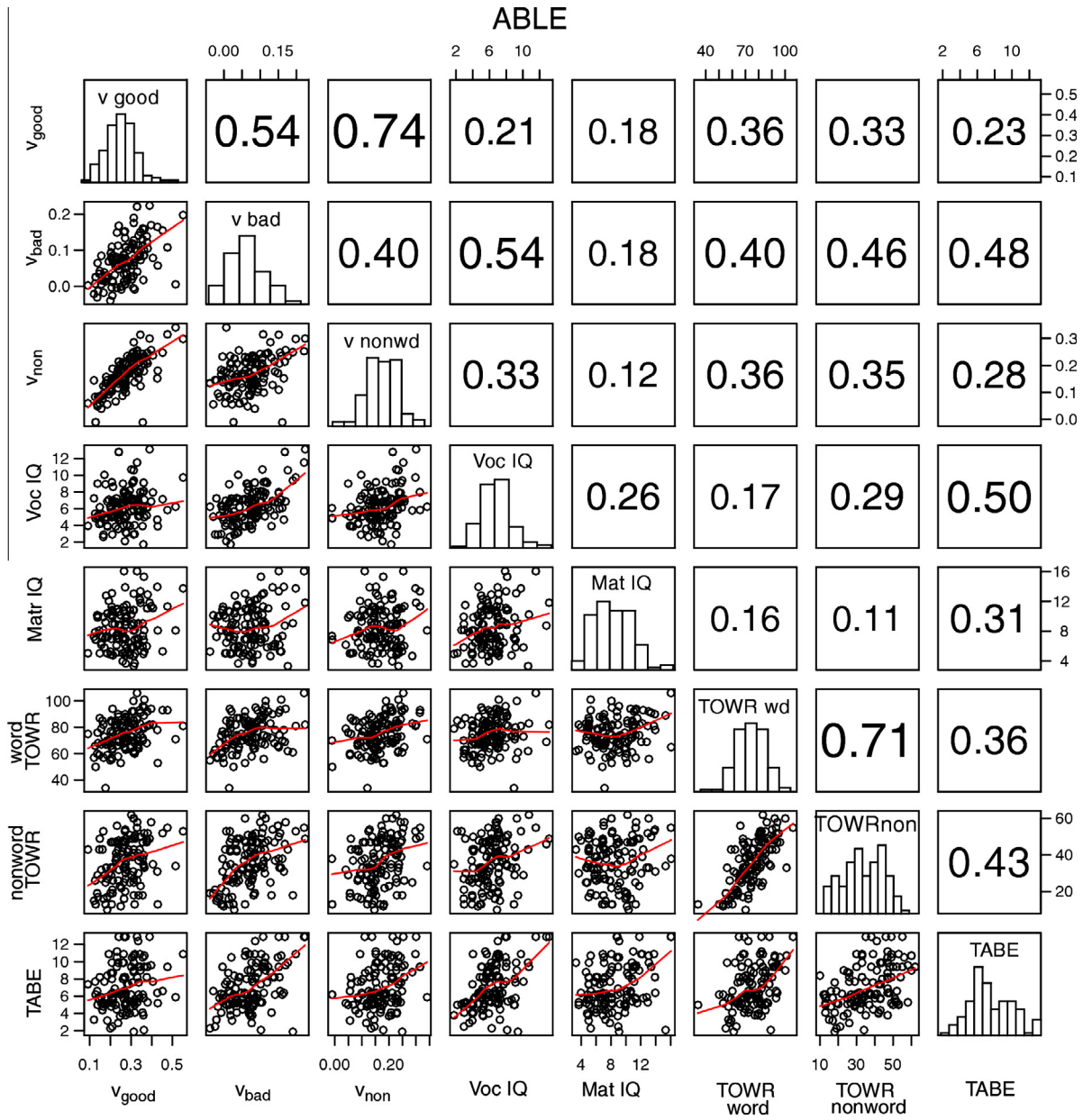

Fig. 4. Scatter plots, histograms, and correlations for drift rates for good and bad words, nonwords, WAIS IQ vocabulary and WAIS matrix reasoning, TOWRE words, TOWRE nonwords, and TABE scores.

settings were significantly positively correlated with accuracy (the slower students were more accurate, the opposite of a speed-accuracy tradeoff). This is not what might be expected if participants who had lower accuracy tried to improve performance by increasing boundary separation, but this may be a motivational issue with undergraduates who participate for credit in a psychology course.

Drift rates for good words correlated with the drift rates for bad words, .70, and with the drift rates for nonwords, .76. The drift rates for bad words correlated .68 with the drift rates for nonwords. Compared with the ABLE students, the correlations in drift rates between good words and nonwords were about the same, but the correlations between bad words and both good words and nonwords were considerably larger. In other words, drift rates lined up the same way for good words, bad words, and nonwords. This contrasts with the ABLE students, for whom scores lined up the same way for good words and nonwords, but they lined up somewhat differently for bad words relative to either good words or nonwords.

7.6. Predictions from combinations of variables

To look at how well combinations of variables predicted a target variable, we used step-wise multiple regression with the TABE scores, WAIS scores, TOWRE scores, and drift rates. The *t*-values for the regression coefficients are shown in Table 4.

7.7. Do combinations of the individual-difference variables predict drift rates?

Drift rates measure the knowledge of words that is represented in lexical memory. In that sense, they may provide an outcome measure that is, in part, the product of an individual's skills with

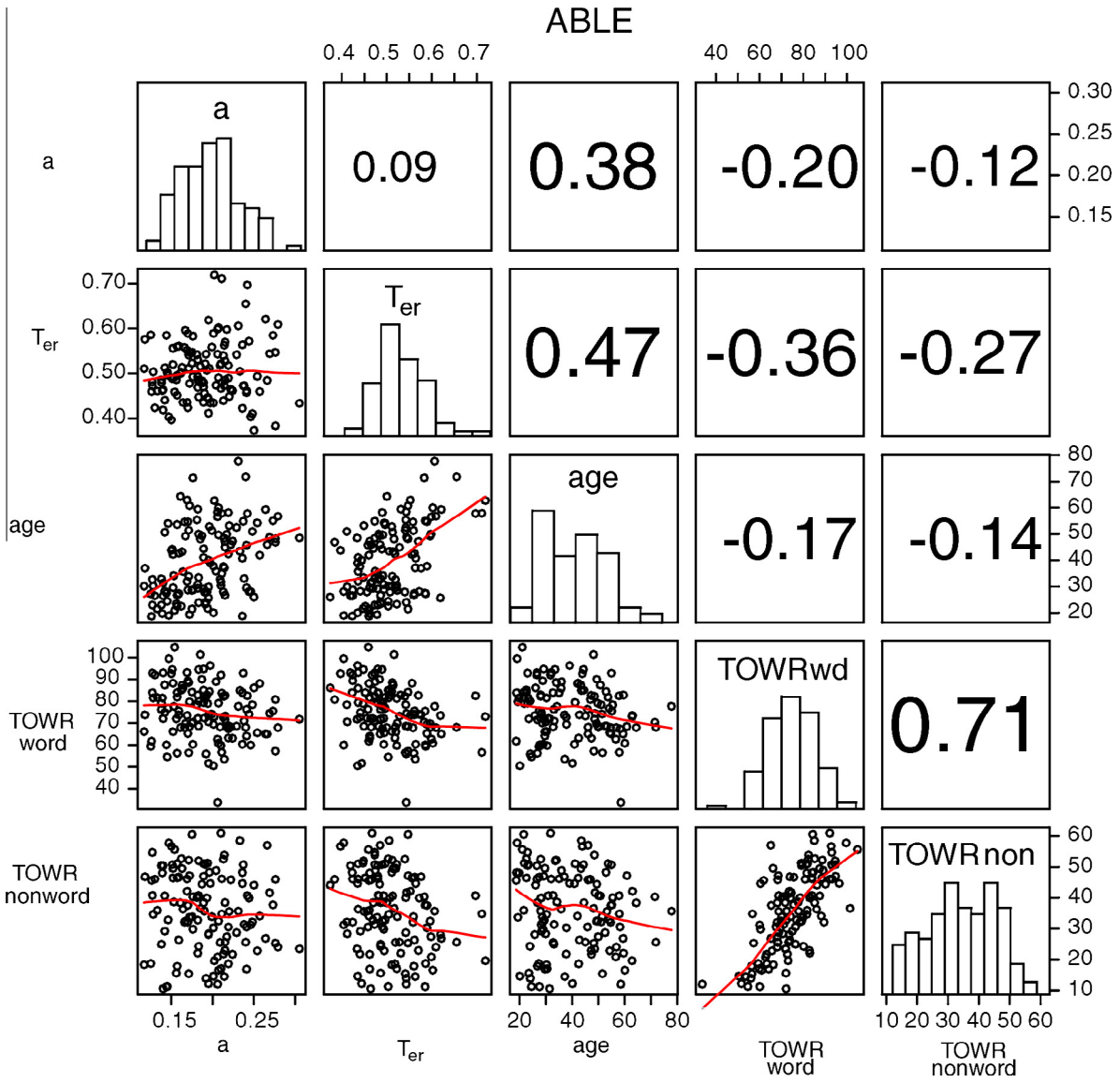

Fig. 5. Scatter plots, histograms, and correlations for boundary separation, nondecision time, age, TOWRE word, and TOWRE nonword scores. Each dot represents an individual participant.

reading words. We performed analyses with the bad words because their drift rates were the most strongly correlated with the individual-difference variables. The TOWRE scores, WAIS scores, drift rates for medium- and low-frequency words, drift rates for nonwords (we took the negative of these so higher numbers mean better performance), TABE scores, and age were entered as predictor variables. The correlations for each factor separately are shown in Fig. 4.

The drift rates for the bad words were predicted by the drift rates for the medium-frequency words and the drift rates for the nonwords; in other words, the bad word drift rates were predicted by how good an individual was at the lexical decision task. The bad word drift rates were also predicted by WAIS vocabulary IQ, which means that there is something more in the drift rates than just how good an individual was at the lexical decision task and this is captured by the vocabulary IQ scores. The results of the multiple regression analyses are shown in Fig. 7, top panel. The conjunction of medium-frequency words' drift rate, nonwords' drift rate, and WAIS vocabulary IQ predicted drift rates with a correlation of .78.

When the drift rates for the medium-frequency words and nonwords were dropped out of the regression, a combination of WAIS vocabulary IQ and TOWRE nonword scores predicted bad words' drift rates (0.64) and when vocabulary IQ was dropped out, TOWRE nonword scores and TABE scores predicted bad words' drift rates (0.56). Thus, overall, there is shared variance between drift rates, vocabulary IQ, TOWRE nonword scores, and TABE scores.

To test reliability and possible over-fitting, we present one representative example of cross-validation for the above combination of WAIS vocabulary IQ and TOWRE nonword scores predicting bad words' drift rates (correlation 0.64). We performed the multiple regression analysis on the data from half of the participants and then examined whether the predictions from the obtained regression coefficients produced a correlation for the other half. The analysis for the first half gave a correlation of 0.67 and when the coefficients were used for the second half, the correlation was 0.58. Thus, the regression coefficients obtained from fits to half the data generalized to the second half of the data.

We performed the same analyses for drift rates for the medium- and low-frequency words and the nonwords as those for the bad words' drift rates. For the medium- and low-frequency words, only the TOWRE word scores significantly predicted drift rates ($r = .38$ and .39, respectively). For the nonwords, TOWRE word scores

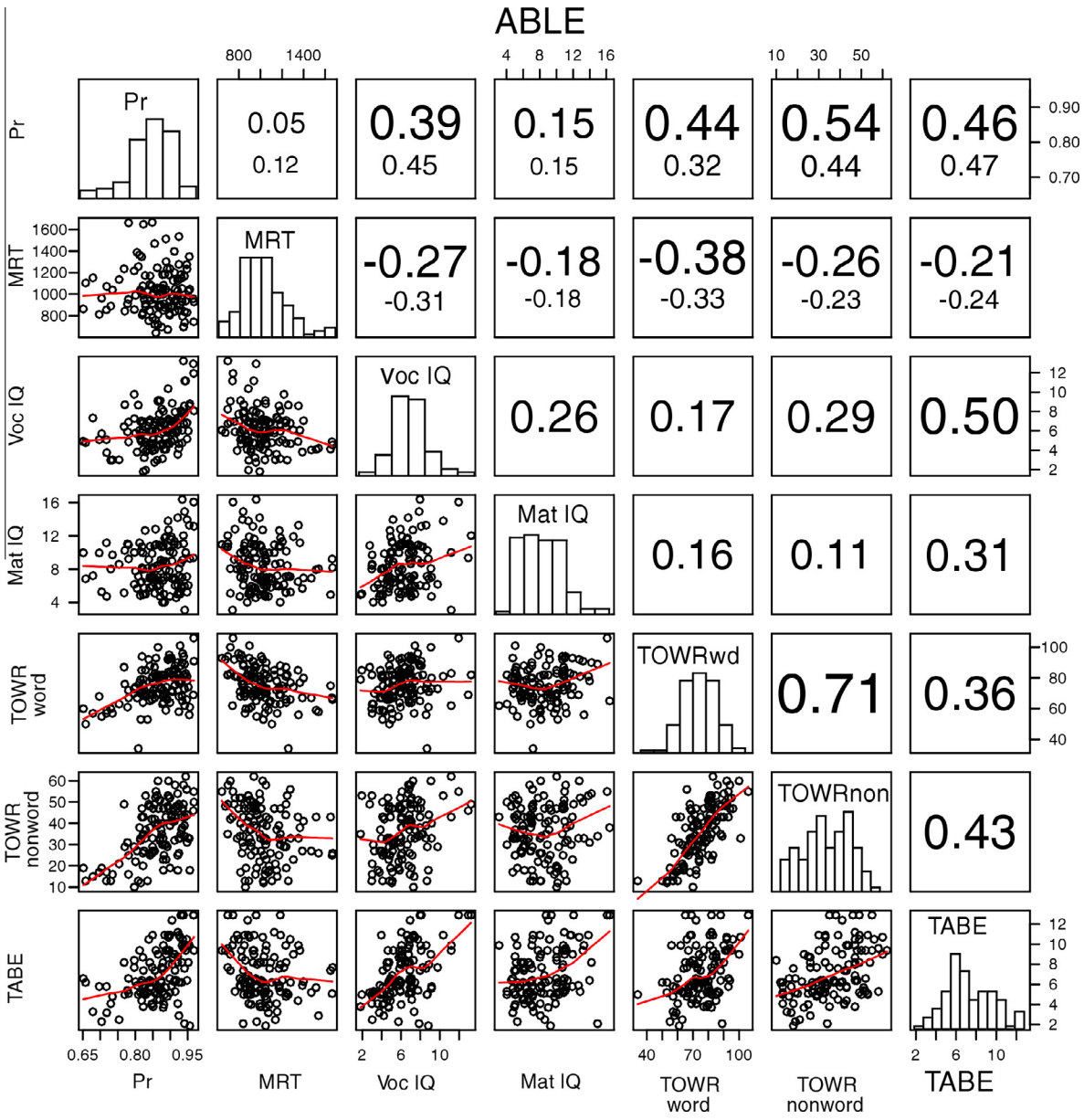

Fig. 6. Scatter plots, histograms, and correlations for mean accuracy (PR) and mean RT across conditions, WAIS vocabulary and WAIS matrix reasoning, TOWRE word, TOWRE nonword, and TABE scores. The second row of correlations for accuracy and mean RT are those for the bad words.

Table 4
Correlations and *t*-values for multiple regressions results.

Predictors	Dependent variable				
	TABE	TABE	Bad word drift	Bad word drift	Bad word drift
TABE	–	–	ns	4.1	ns
Bad word drift	ns	3.9	–	–	–
Med. freq. drift	ns	ns	–	–	6.7
Nonword drift	ns	ns	–	–	2.7
WAIS vocab.	3.4	–	6.1	–	5.5
WAIS matrix	2.3	2.9	ns	ns	ns
TOWR word	ns	ns	ns	ns	ns
TOWR nonword	3.0	3.0	2.1	3.7	ns
R	.63	.58	.64	.56	.78

Note: A dash means the variables was not entered in the multiple regression and ns means the variable was not significant.

and WAIS vocabulary IQ predicted drift rates ($r = .45$). None of these correlations were as strong as those for the drift rates for the bad words.

7.8. Do combinations of the individual-difference variables predict TABE scores?

TABE scores are of importance from a practical perspective because they are the basis for ABLE class placement. We performed step-wise multiple regression with the scores on the TABE as the dependent variable. WAIS vocabulary IQ, WAIS matrix reasoning IQ, and TOWRE nonword scores predicted TABE scores with a correlation of .63, but the drift rates for good words, bad words, and nonwords were not significant predictors; they explained no variance in TABE scores beyond that explained by vocabulary IQ and TOWRE nonword scores.

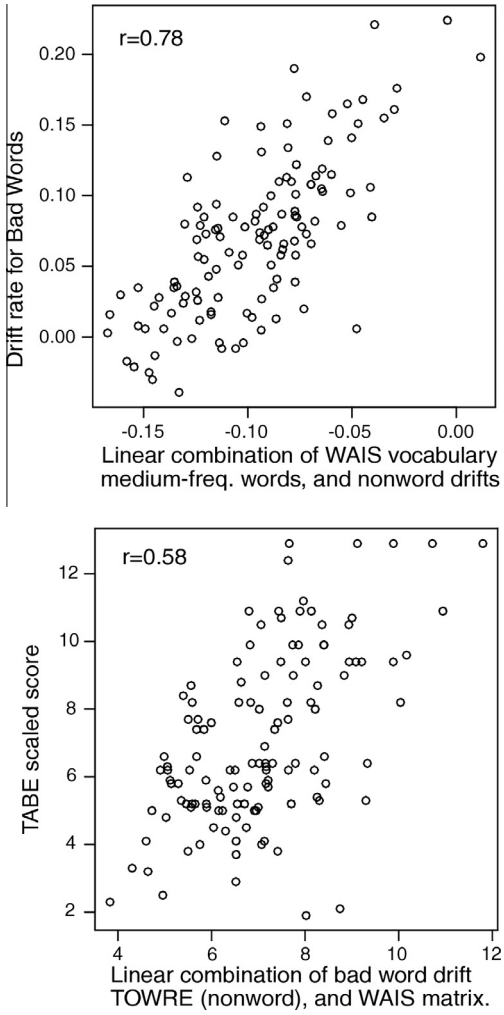

Fig. 7. The top panel shows drift rates for bad words plotted against the predictor values from step-wise multiple regression with WAIS vocabulary IQ, TOWRE nonword scores, and age as predictors. The bottom panel shows scaled TABE scores plotted against bad word drift rates, TOWRE nonword scores, and WAIS matrix reasoning scores as predictors.

To show that vocabulary IQ and bad-word drift rates served the same purpose in the regression analysis, we dropped out vocabulary IQ. The correlation was reduced modestly to .58 and drift rates for the bad words replaced vocabulary IQ as a predictor (predictions are shown in Fig. 7, bottom panel), demonstrating that drift rates for the bad words were related to scores on the TABE placement test. The TABE involves answering questions about just-read paragraphs, which means that it reflects skills that build representations of multiple pieces of information, with appropriate connections among the pieces and appropriate inferences from them. Also, as described earlier, a TABE score could come from one of four different tests (the scores are then scaled). It is somewhat surprising that, despite this complexity, the drift rates for the bad words were still a significant predictor.

7.9. Are the bad words bad simply because they are low-frequency words?

It might be hypothesized that there is nothing special about our bad words, that all of the findings for them can be accounted for by their word frequency. The low-frequency words in the experiment (data and drift rates presented in Tables 2 and 3) had a mean Kucera-Francis word frequency of 9.23 whereas the bad words had a mean frequency of 4.99, so it could be that this is the reason for their high correlations with the TOWRE and TABE tests and their high predictive values in the multiple regression analyses. However, this was not the case. To show this, we divided the low-frequency words into two subsets, one with frequencies higher than 10 and one with frequencies between 1 and 10. The frequencies of the latter group had a mean of 4.89, which matches the bad words. We fit the diffusion model to the data as before but with the low-frequency words divided into the two subsets (giving 5 drift rates instead of the 4 shown in Table 3).

We found that the drift rates for the low-frequency words that matched the bad words had much lower correlations with all of the measures (WAIS, TOWRE, TABE) than the drift rates for the bad words. For example, we examined which variables predicted drift rates for these matched low-frequency words (without the other drift rates), as in the third column of Table 4. The significant predictors were WAIS vocabulary and TOWRE words, with a correlation of .35. For comparison, for bad words the predictors were WAIS vocabulary and TOWRE nonwords, with a correlation of .64. We also examined whether the drift rates for the matched low-frequency words were significant in accounting for TABE scores when WAIS vocabulary scores were not included in the multiple regression, as was the case for the bad word drift rates, but they were not. From these comparisons, we conclude that there is something special for our population of low-literacy adults beyond word frequency for our set of bad words.

8. Discussion

The results from this study show the potential benefits of conducting model-based analyses of psycholinguistic tasks in combination with standardized IQ, achievement, and reading tests. Lexical decision is a task for which responses are fast, passive, and automatic and it has been repeatedly demonstrated in psycholinguistic research to assess the knowledge of words that is used in reading, the knowledge that is abstracted by the diffusion model from RTs and accuracy (Ratcliff, Gomez et al., 2004). Because the standard deviations in the model's parameter values were less than the standard deviations among participants, meaningful correlations could be computed among the mechanisms identified by the model and individual-difference variables. The significant correlations between drift rates and WAIS vocabulary IQ, TOWRE scores, and TABE scores verify that the lexical decision task measures information that has been considered important for discriminating the skills with which readers do and do not have difficulty. To put it another way, lexical decision is a task known to assess lexical knowledge and so when the diffusion model abstracts drift rates that correlate with the IQ, TOWRE, and TABE scores, it maps those scores to lexical knowledge. It is also the case that correlations among drift rates and the scores were higher for the bad words in our study, indicating that they reflected near the limit of the ABLE students' knowledge.

The structure of the diffusion model separates components of processing from each other. In support of this, there were no significant correlations among boundary settings, nondecision times, or drift rates. It would seem possible that individuals like the ABLE students who have difficulties in reading words (i.e., low drift rates) would attempt to counter these problems by setting their boundaries further apart than individuals with better skills, but this did not happen. For the TOWRE tests in particular, it might have been expected that the scores would be significantly correlated with boundaries because they are timed tests. Likewise, it would seem possible that an individual who is limited in the speed with which he or she can carry out nondecision processes (e.g., encoding strings of letters, accessing lexical memory, executing

responses) might set their boundaries further apart, but again this did not happen. It may be that an individual's boundary settings are to some large degree determined by his or her decision-making style across all of the many decision-making domains of everyday life.

A result we stress is that speed and accuracy were not significantly correlated across the participants in the study (as they have not been in previous studies with a variety of tasks, e.g., Ratcliff et al., 2010, 2015). A participant's accuracy did not imply anything about his or her speed and a participant's speed did not imply anything about his or her accuracy. It follows from this that individuals cannot be equated on the basis of accuracy alone. Two individuals (or two groups of individuals) who have the same level of accuracy cannot be said to have the same skills because their speeds are likely to be different. The faulty conclusions that can be drawn when using only one of the dependent variables have been demonstrated with the finding that older adults' responses are not slower than young adults' because their drift rates are lower but rather because they set their boundaries farther apart and their nondecision times are longer (Ratcliff et al., 2001, 2003, 2010, 2011; Ratcliff, Gomez et al., 2004). This issue is relevant to interpretations of scores on the TOWRE, TABE, and IQ tests because they were all significantly correlated with RTs as well as accuracy. For the untimed tests especially (the TABE and IQ tests), it might have been thought that accuracy is directly related to comprehension and knowledge of words and this is the implicit assumption that underlies most uses of the tests. But the findings that the tests also reflect speed and that speed and accuracy are independent means that speed and accuracy must both be considered.

There is a puzzle in the individual differences literature as to why RTs in general and lexical decision RTs in particular are related to some but not all reading measures that involve speeded processing. For example, Katz et al. (2011) examined RTs for lexical decision and naming for 99 adults with a wide range of reading abilities and found that RTs on lexical decision and naming correlated strongly with word identification (as measured by subtests of the Woodcock–Johnson, TOWRE, and GORT reading tests), modestly with vocabulary size (as measured by subtests of the Woodcock–Johnson, the Wechsler Abbreviated Scale of Intelligence, and the Peabody Picture Vocabulary Test), and not at all with reading comprehension (as measured by the Woodcock–Johnson and GORT comprehension tests). Accuracy for lexical decision was high, around 90% correct, and was not used in any of the analyses.

In another example, Stringer and Stanovich (2000) examined RTs for two-choice tasks that used simple visual stimuli (choosing between a circle with a line through it and three squares) and simple auditory stimuli (choosing between two tones of different frequencies). The participants were 81 adults with a wide range of reading abilities. Stringer and Stanovitch found that there was little direct relationship between the RTs and phonological awareness, general cognitive ability, or word recognition ability.

Our results from application of the diffusion model suggest a way to understand the results of these two studies. Because individual differences in drift rate are determined mainly by accuracy and individual differences in RT are determined mainly by boundary separation and nondecision time and there are no correlations between accuracy and RT, the studies may have picked the wrong dependent measure to use. However, because accuracy was high in those studies (near ceiling), individual differences may not have been particularly strong. This reflects the larger issue that in the domain of word comprehension, the primary dependent variable is usually RT. As we have shown here, RTs are less useful for examining individual differences (they produce weaker relationships with reading scores) and the diffusion model analysis provides a way of understanding why this occurs.

This leads into a methodological point concerning whether drift rates or accuracy values provide the best dependent variable. Our results showed that drift rates produced marginally higher correlations than accuracy values. In other studies, drift rates have tracked accuracy and the two provide similar interpretations of individual differences. However, if accuracy approaches ceiling, as it might for high frequency words in lexical decision, drift rates become more constrained by RTs than accuracy (Ratcliff, 2014). This means that power is increased for drift rates relative to accuracy values. Also, if there are relatively few critical items in a condition of an experiment (e.g., clinical applications), then the drift rates obtained from fitting the data for the critical items and filler items simultaneously can produce twice the power to detect differences as accuracy or RTs (McKoon & Ratcliff, 2012; White et al., 2010a).

We note a previous application (Naples, Katz, & Grigorenko, 2012) of the diffusion model to individual differences in several reading-related studies (the participants were Russian children with ages between 7 and 12). Naples et al. measured RTs and accuracy for a task in which the children were asked to decide whether two strings of letters or two strings of numbers matched. They found that the reading-related measures correlated strongly with drift rates but only weakly with nondecision times. Their conclusion was that speed of processing, as measured by RTs, is related to reading but mainly expressed through drift rates (see also Zeguers et al., 2011). This is the same result that we found, that RTs were less strongly correlated than drift rates with reading measures.

9. Conclusions

For low-literacy individuals, the study reported here is the first to begin to break reading comprehension into the mechanisms responsible for it in such a way as to separate the information on which responses are based from the speed with which they are made. This separation allows exploration of relations between individual-difference measures and underlying lexical knowledge and it allows exploration of relations among individual-difference measures in terms of underlying knowledge, rather than directly to each other.

Our results point to the need for future research that attempts to better understand what it is that leads to significant correlations among some tests and processing mechanisms but not others. For example, it might be fruitful to test readers on lexical decision, one or more of the individual-difference measures we used, and other standardized measures of reading difficulties such as the Woodcock–Johnson tests.

WAIS vocabulary IQ was significantly correlated with scores on all three standardized tests that we used (the TOWRE and TABE tests) and it was significantly correlated with the knowledge of words that is measured by drift rates in the diffusion model. The strength of the correlations suggests that tests of WAIS vocabulary IQ should be regularly used in conjunction with other standardized tests and with new tests that may be developed. In addition, WAIS vocabulary IQ may be especially important in designing interventions to improve reading skills because different levels of IQ may require different kinds of interventions. From a theoretical perspective, the words used in the WAIS vocabulary test increase in difficulty through the test, but they seem to tap into the same assessment of difficulty that our bad words (sorted on lexical decision performance) did.

It is important to point out that despite the more strategic nature of the TABE, we find that drift rates for lexical decision are quite highly correlated with the TABE. So an important component of the TABE is captured by word knowledge, especially that part of

knowledge related to words that are at the limit of knowledge (our "bad" words) for this class of participants.

To summarize, it is our hope that, for low-literacy individuals, cognitive psychology can provide methods that can successfully investigate fast automatic processes, real-time reading processes, and speed/accuracy tradeoffs, and their consequences. Standardized reading tests like the TOWRE and TABE tests can get us only so far. They can show relationships between some aspects of processing and reading, for example, phonological coding, knowledge of vocabulary, and so on, but they themselves involve many different mechanisms of reading. The TOWRE words test, for example, may involve knowledge about the regularity and frequency of occurrence of letter and phonemic bigrams, trigrams, and syllable structures, and the meanings of words. It also may involve knowledge about the mappings between variables like those just listed and the processes that produce spoken words. It has a large speed component that might be closely intertwined with the variables or that might have some aspects that are separate and malleable. Because tests like the TOWRE tests and the TABE have so many different aspects, they do not pinpoint specific processes in the way that cognitive models sometimes can.

Acknowledgments

This work was supported by the Institute for Educational Sciences (grant number R305A120189) and the National Institute on Aging (grant number R01-AG041176). We thank Daphne Greenberg for comments on this article.

References

Baer, J., Kutner, M., & Sabatini, J. (2009). *Basic reading skills and the literacy of America's least literate adults: Results from the 2003 National Assessment of Adult Literacy (NAAL) supplemental studies (NCES 2009-481)*. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Beder, H. (1999). *The outcomes and impacts of adult literacy education in the United States*. Cambridge, MA: National Center for the Study of Adult Learning and Literacy.

Calhoon, M. B., Scarborough, H. S., & Miller, B. (2013). Interventions for struggling adolescent and adult readers: Instructional, learner, and situational differences. *Reading and Writing, 26*, 489–494.

Donkin, C., Brown, S., Heathcote, A., & Wagenmakers, E. J. (2011). Diffusion versus linear ballistic accumulation: Different models for response time, same conclusions about psychological mechanisms? *Psychonomic Bulletin & Review, 55*, 140–151.

Dufau, S., Grainger, J., & Ziegler, J. C. (2012). How to say "no" to a nonword: A leaky competing accumulator model of lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 1117–1128.

Ehringhaus, C. (1991). Teachers' perceptions of testing in adult basic education. *Adult Basic Education, 1*, 138–155.

Geddes, J., Ratcliff, R., Allerhand, M., Childers, R., Wright, R. J., Frier, B. M., & Deary, I. J. (2010). Modeling the effects of hypoglycemia on a two-choice task in adult humans. *Neuropsychology, 24*, 652–660.

Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review, 103*, 518–565.

Greenberg, D. (2008). The challenges facing adult literacy programs. *Community Literacy Journal, 3*, 39–54.

Harm, M. W., McCandliss, B. D., & Seidenberg, M. S. (2003). Modeling the successes and failures of interventions for disabled readers. *Scientific Studies of Reading, 7*, 155–182.

International Adult Literacy Survey Institute (2011). Retrieved from the website: <http://www.statcan.gc.ca/pub/89-588-x/4152887-eng.htm> 03.03.13.

Jones, M., & Dzhafarov, E. N. (2014). Unfalsifiability and mutual translatability of major modeling schemes for choice reaction time. *Psychological Review, 121*, 1–32.

Katz, L., Brancazio, L., Irwin, J., Katz, S., Magnuson, J., & Whalen, D. H. (2011). What lexical decision and naming tell us about reading. *Reading and Writing, 25*, 1259–1282.

Keenan, J. M., & Meenan, C. E. (2014). Test differences in diagnosing reading comprehension deficits. *Journal of Learning Disabilities, 47*, 125–135.

Kristof, N. (2014). The American dream is leaving America. Retrieved from the website: <http://www.nytimes.com/2014/10/26/opinion/sunday/nicholas-kristof-the-american-dream-is-leaving-america.html> 22.07.15.

Kutner, M., Greenberg, E., & Baer, J. (2006). *National Assessment of Adult Literacy (NAAL): A first look at the literacy of America's adults in the 21st century (Report No. NCES 2006-470)*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

MacArthur, C. A., Konold, T. R., Glutting, J. J., & Alamprese, J. A. (2010). Reading component skills of learners in adult basic education. *Journal of Learning Disabilities, 43*, 108–121.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review, 88*, 375–407.

McKoon, G., & Ratcliff, R. (2012). Aging and IQ effects on associative recognition and priming in item recognition. *Journal of Memory and Language, 66*, 416–437.

McKoon, G., & Ratcliff, R. (2015). Cognitive theories in discourse-processing research. In E. J. O'Brien, A. E. Cook, & R. F. Lorch, Jr. (Eds.), *Inferences during reading*. Cambridge: Cambridge University Press.

McLeod, P., Shallice, T., & Plaut, D. C. (2000). Visual and semantic influences in word recognition: Converging evidence from acquired dyslexic patients, normal subjects, and a computational model. *Cognition, 74*, 91–114.

Mellard, D. F., Fall, E., & Woods, K. L. (2010). A path analysis of reading comprehension for adults with low literacy. *Journal of Learning Disabilities, 43*, 154–165.

Mellard, D. F., Woods, K. L., Desa, Z. D. M., & Vuyk, M. A. (2013). Underlying reading-related skills and abilities among adult learners. *Journal of Learning Disabilities*. http://dx.doi.org/10.1177/0022219413500813 [first published on August 20, 2013 as].

Miller, B., McCardle, P., & Hernandez, R. (2010). Advances and remaining challenges in adult literacy research. *Journal of Learning Disabilities, 43*, 101–107.

Mulder, M. J., Bos, D., Weusten, J. M. H., van Belle, J., van Dijk, S. C., Simen, P., ... Durson, S. (2010). Basic impairments in regulating the speed-accuracy tradeoff predict symptoms of attention-deficit/hyperactivity disorder. *Biological Psychiatry, 68*, 1114–1119.

Nanda, A. O., Greenberg, D., & Morris, R. (2010). Modeling child-based theoretical reading constructs with struggling adult readers. *Journal of Learning Disabilities, 43*, 139–153.

Naples, A., Katz, L., & Grigorenko, E. L. (2012). Reading and a diffusion model analysis of reaction time. *Developmental Neuropsychology, 37*, 299–316.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited capacity intention. *Journal of Experimental Psychology: General, 106*, 226–254.

Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review, 113*, 327–357.

OECD (2013). *Time for the U.S. to Reskill?: What the survey of adult skills says, OECD skills studies*. OECD Publishing. http://dx.doi.org/10.1787/9789264204904-en.

Patterson, K., Seidenberg, M. S., & McClelland, J. L. (1989). Connections and disconnections: Acquired dyslexia in a computational model of reading processes. In R. Morris (Ed.), *Parallel distributed processing: Implications for psychology and neurobiology*. New York: Oxford University Press.

Perry, C., Ziegler, J. C., & Zorzi, M. (2007). Nested incremental modeling in the development of computational theories: The CDP+ model of reading aloud. *Psychological Review, 114*, 273–315.

Posner, M. I., & Snyder, C. R. (1975b). Facilitation and inhibition in the processing of signals. In P. M. A. Rabbitt (Ed.), *Attention and performance V*. London: Academic Press.

Posner, M. I., & Snyder, C. R. (1975a). Attention and cognitive control. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola symposium* (pp. 55–85). Hillsdale, N.J.: Erlbaum.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*, 59–108.

Ratcliff, R. (2002). A diffusion model account of reaction time and accuracy in a two choice brightness discrimination task: Fitting real data and failing to fit fake but plausible data. *Psychonomic Bulletin and Review, 9*, 278–291.

Ratcliff, R. (2014). Measuring psychometric functions with the diffusion model. *Journal of Experimental Psychology: Human Perception and Performance, 40*, 870–888.

Ratcliff, R., & Childers, R. (2015). Individual differences and fitting methods for the two-choice diffusion model of decision making. *Decision, 2*, 237–279.

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A diffusion model account of the lexical-decision task. *Psychological Review, 111*, 159–182.

Ratcliff, R., Love, J., Thompson, C. A., & Opfer, J. (2012). Children are not like older adults: A diffusion model analysis of developmental changes in speeded responses. *Child Development, 83*, 367–381.

Ratcliff, R., & McKoon, G. (1981). Automatic and strategic priming in recognition. *Journal of Verbal Learning and Verbal Behavior, 20*, 204–215.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20*, 873–922.

Ratcliff, R., Perea, M., Colangelo, A., & Buchanan, L. (2004). A diffusion model account of normal and impaired readers. *Brain & Cognition, 55*, 374–382.

Ratcliff, R., Schmiedek, F., & McKoon, G. (2008). A diffusion model explanation of the worst performance rule for reaction time and IQ. *Intelligence, 36*, 10–17.

Ratcliff, R., Thapar, A., Gomez, P., & McKoon, G. (2004). A diffusion model analysis of the effects of aging in the lexical-decision task. *Psychology and Aging, 19*, 278–289.

Ratcliff, R., Thapar, A., & McKoon, G. (2001). The effects of aging on reaction time in a signal detection task. *Psychology and Aging, 16*, 323–341.

Ratcliff, R., Thapar, A., & McKoon, G. (2003). A diffusion model analysis of the effects of aging on brightness discrimination. *Perception and Psychophysics, 65*, 523–535.

Ratcliff, R., Thapar, A., & McKoon, G. (2006). Aging and individual differences in rapid two-choice decisions. *Psychonomic Bulletin and Review, 13*, 626–635.

Ratcliff, R., Thapar, A., & McKoon, G. (2007). Application of the diffusion model to two-choice tasks for adults 75–90 years old. *Psychology and Aging, 22*, 56–66.
Ratcliff, R., Thapar, A., & McKoon, G. (2010). Individual differences, aging, and IQ in two-choice tasks. *Cognitive Psychology, 60*, 127–157.
Ratcliff, R., Thapar, A., & McKoon, G. (2011). Effects of aging and IQ on item and associative memory. *Journal of Experimental Psychology: General, 140*, 46–487.
Ratcliff, R., Thapar, A., Smith, P. L., & McKoon, G. (2005). Aging and response times: A comparison of sequential sampling models. In J. Duncan, P. McLeod, & L. Phillips (Eds.), *Speed, control, and age*. Oxford, England: Oxford University Press.
Ratcliff, R., Thompson, C. A., & McKoon, G. (2015). Modeling individual differences in response time and accuracy in numeracy. *Cognition, 137*, 115–136.
Ratcliff, R., & Tuerlinckx, F. (2002). Estimating the parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin and Review, 9*, 438–481.
Ratcliff, R., & Van Dongen, H. P. A. (2009). Sleep deprivation affects multiple distinct cognitive processes. *Psychonomic Bulletin and Review, 16*, 742–751.
Schmiedek, F., Oberauer, K., Wilhelm, O., Suβ, H.-M., & Wittmann, W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *Journal of Experimental Psychology: General, 136*, 414–429.
Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96*, 523–568.
Smith, P. L., Ratcliff, R., & McKoon, G. (2014). The diffusion model is not a deterministic growth model: Comment on Jones and Dzhafarov (2014). *Psychological Review, 121*, 679–688.
Spaniol, J., Madden, D. J., & Voss, A. (2006). A diffusion model analysis of adult age differences in episodic and semantic long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 101–117.
Starns, J. J., & Ratcliff, R. (2010). The effects of aging on the speed-accuracy compromise: Boundary optimality in the diffusion model. *Psychology and Aging, 5*, 377–390.
Stringer, R., & Stanovich, K. E. (2000). The connection between reaction time and variation in reading ability: Unravelling covariance relationships with cognitive ability and phonological sensitivity. *Scientific Studies of Reading, 4*, 41–53.
Tighe, E. L., & Schatschneider, C. (2014). Examining the relationships of component reading skills to reading comprehension in struggling adult readers: A meta-analysis. *Journal of Learning Disabilities*. http://dx.doi.org/10.1177/0022219414555415 [published online before print October 16, 2014].
Torgesen, J. K., & Wagner, R. (1999). *Test of word reading efficiency*. Austin, TX: Pro-Ed.
Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review, 108*, 550–592.
Vandekerckhove, J., & Tuerlinckx, F. (2008). Diffusion model analysis with MATLAB: A DMAT primer. *Behavior Research Methods, 40*, 61–72.
Voss, A., & Voss, J. (2007). Fast-dm: A free program for efficient diffusion model analysis. *Behavior Research Methods, 39*, 767–775.
Wagenmakers, E.-J., Van Der Maas, H. L. J., & Grassman, R. P. P. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin and Review, 14*, 3–22.
White, C., Ratcliff, R., Vasey, M., & McKoon, G. (2009). Dysphoria and memory for emotional material: A diffusion model analysis. *Cognition and Emotion, 23*, 181–205.
White, C. N., Ratcliff, R., Vasey, M. W., & McKoon, G. (2010a). Using diffusion models to understand clinical disorders. *Journal of Mathematical Psychology, 54*, 39–52.
White, C. N., Ratcliff, R., Vasey, M. W., & McKoon, G. (2010b). Anxiety enhances threat processing without competition among multiple inputs: A diffusion model analysis. *Emotion, 10*, 662–677.
Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Frontiers in Neuroinformatics, 7*, 1–10.
Zeguers, M. H. T., Snellings, P., Tijms, J., Weeda, W. D., Tamboer, P., Bexkens, A., & Huizenga, H. M. (2011). Specifying theories of developmental dyslexia: A diffusion model analysis of word recognition. *Developmental Science, 14*, 1340–1354.
Ziegler, J. C., Castel, C., Pech-Georgel, C., George, F., Alario, F.-X., & Perry, C. (2008). Developmental dyslexia and the dual route model of reading: Simulating individual differences and subtypes. *Cognition, 107*, 151–178.