SCIENTIFIC STUDIES OF READING
2017, VOL. 21, NO. 4, 292–309
https://doi.org/10.1080/10888438.2017.1287188

Routledge
Taylor & Francis Group

Adults With Poor Reading Skills and the Inferences They Make During Reading

Gail McKoon and Roger Ratcliff

The Ohio State University

ABSTRACT

Millions of U.S. adults lack the literacy skills needed for most living-wage jobs. We investigated one particular comprehension process for these adults: generating predictive inferences. If a sentence says that someone falls from a 14th-story roof, a reader should infer almost certain death. On any test of comprehension, there are two dependent variables: the speed of the response to a test item and accuracy. To address both simultaneously, we used a decision model that separates how much information an individual understands from a text and the individual's speed/accuracy trade-off settings. We found that adult literacy students do differentiate between predictive inference sentences and control sentences, a finding that illustrates how a decision-making model combined with tests of particular comprehension processes can lead to further understanding of low-literacy adults' reading skills.

The number of adults in the United States who have only the lowest of literacy skills is staggeringly high (Baer, Kutner, Sabatini, & White, 2009; Greenberg, 2008; Kutner, Greenberg, & Baer, 2006; Miller, McCardle, & Hernandez, 2010; OECD, 2013). As Nicholas Kristof (2014) of the *New York Times* put it recently, these data "should be a shock to Americans." In this article, we bring theory and methodology from cognitive psychology studies of reading to bear on one particular aspect of this problem—the extent to which low-literacy adults make inferences that are necessary to achieve full comprehension of textual information.

We place the research described in this article in general theories of reading comprehension. Perfetti and Stafura (2015; also Cromley & Azevedo, 2007), for example, laid out the many processes and sorts of information involved in reading and the connections that tie them together. Inferences play an important role in this framework. They take the explicit information that is put together from the words of a text and construct out of them a representation of the meaning of the text that goes beyond what is present in its individual words. There are many kinds of inferences, ranging from relatively simple ones like connecting a pronoun to its referent to more complex ones like understanding causal relationships. For some inferences, it has been suggested that poor readers have difficulties; they may not appropriately integrate pieces of information in a text with each other, integrate pieces of information with general knowledge, establish appropriate causal connections among pieces of information, interpret textual information in a context-sensitive way, identify the main ideas of a text, or establish the referents of pronouns, and it has been suggested that difficulties like these can occur even when a reader has the requisite general knowledge upon which to base an inference (Barnes, Ahmed, Barth, & Francis, 2015; Bowyer-Crane & Snowling, 2005; Cain & Oakhill, 1999, 2006; Cain, Oakhill, Barnes, & Bryant, 2001; Cain, Oakhill, & Lemmon, 2004; Garnham, Oakhill, & Johnson-Laird, 1982; Laing & Kamhi, 2003; Long & Golding, 1993; Long, Oppy, & Seely, 1994; Magliano & Millis, 2003; Oakhill, 1982, 1983, 1984, 1993; Oakhill & Yuill, 1986; Oakhill, Yuill, &

CONTACT Gail McKoon mckoon.1@osu.edu Department of Psychology, The Ohio State University, 1835 Neil Avenue, Columbus, OH, 43210.

Donaldson, 1990; Oakhill, Yuill, & Parkin, 1988; Singer, Andrusiak, Reisdorf, & Black, 1992; Todaro, Millis, & Dandotkar, 2010; Whitney, Ritchie, & Clark, 1991; Yuill, Oakhill, & Parkin, 1989).

Among inferences that have been little investigated are predictive inferences. These are inferences about what will happen next in an episode or story, as death will almost certainly happen when an actress falls from the 14th story of a building.[1] The first research on predictive inferences was done by McKoon and Ratcliff (1986, 1992). Their study was pivotal in that it moved the field to make a distinction between fast, automatic, passive reading processes and slower, conscious, more strategic ones. Processes of the former kind support much of the information that is understood during reading, with strategic processes coming into play when a reader has specific goals for comprehension (such as a full and complete understanding of a research article). The 1986 article led to a theoretical framework in which fast, automatic, passive processes are based on information that is easily available from memory, either general knowledge or information from a text that is currently being read (McKoon & Ratcliff, 1992; O'Brien, 1995).

This memory-based processing framework is now the most widely accepted view of fast automatic comprehension processes in cognitive psychology (e.g., Cook, Limber, & O'Brien, 2001; Cook & O'Brien, 2015; Gerrig & McKoon, 1998; Gerrig & O'Brien, 2005; McKoon & Ratcliff, 1998, 2015). These processes are now often labeled "resonance," a term first coined by Lockhart, Craik, and Jacoby (1976) and taken up by O'Brien and colleagues (O'Brien & Myers, 1999; see O'Brien, Cook, & Lorch, 2015, for a review). There have been a number of hypotheses about how automatic processes carry through a text. In van den Broek's (e.g., van den Broek, Beker, & Oudega, 2015) landscape view, automatic processes make available new and different information over the course of a text. Cook and O'Brien (2015) and Isberner and Richter (2014) suggested three processing stages, with the first making information from memory available by automatic processes, the second integrating information from memory with information in working memory, and the third validating that information against information in memory. In constructionist hypotheses (e.g., Graesser, Li, & Feng, 2015), automatic processes support the strategic construction of inferences about, for example, causality. Many of these researchers assume the scheme developed by Kintsch and colleagues (e.g., van Dijk & Kintsch, 1983) in which automatic processes make available information from memory and further automatic processes integrate that information with information in working memory.

Given the pivotal role of predictive inferences in highlighting the distinction between automatic processes and strategic ones, it is surprising that they have not been investigated thoroughly in research with low-literacy readers. To begin such an effort, the study reported here used sentences like, "The director and cameraman were ready to shoot close-ups when suddenly the actress fell from the 14th story" (McKoon & Ratcliff, 1986). College students encode this kind of inference quickly and automatically during reading (e.g., Casteel, 2007; Cook et al., 2001; Gueraud, Tapiero, & O'Brien, 2008; Lassonde & O'Brien, 2009; Linderholm & van den Broek, 2002; Murray & Burke, 2003; Peracchi & O'Brien, 2004; Sanford & Garrod, 2005; see Gerrig & O'Brien, 2005; O'Brien et al., 2015; and van den Broek, Rapp, & Kendeou, 2005, for reviews). The question was whether low-literacy readers also do so. To preview the results of our study, we found that they do, in contrast to the many studies just listed in which it appears that they do not make appropriate inferences. In the Discussion section, we discuss our result, why it might differ from results for other types of inferences, and how it might fit into an overall picture of low-literacy adults' inferential abilities.

An important difference between our study and those just cited is that we explicitly consider the speed/accuracy trade-offs that individuals can adopt. Understanding these trade-offs is essential. An individual may respond with low accuracy to test items because he or she does not know the relevant information or because he or she does know the information but decides to prioritize speed over accuracy. This means that the information that an individual has cannot be determined from speed

[1]Death is not an absolutely necessary outcome. A 29-year-old window-washer fell 47 floors from an apartment building and survived, as did a 29-year-old man who plunged 17 stories in the atrium of a hotel in Minneapolis, a 22-year-old amateur skydiver who went into free fall more than a mile above the earth (McFadden, 2007), and a cat that fell 19 stories from a high-rise window (Hager, 2012).

alone or from accuracy alone. Of relevance here, it cannot be determined whether a reader does or does not encode predictive inferences. Another way to say this point is that individuals who perform with the same speed may have differences in accuracy, and therefore differences in the information underlying their performance, and individuals with the same accuracy may have differences in speed, and therefore differences in the information underlying their performance. Measuring accuracy alone or response times (RTs) alone will almost certainly give misleading interpretations of data (e.g., Ratcliff, Thapar, & McKoon, 2010, 2011; Ratcliff, Thompson, & McKoon, 2015).

To handle this problem, we use a decision-making model (Ratcliff's diffusion model; Ratcliff, 1978; Ratcliff & McKoon, 2008) that translates accuracy and RTs into the same underlying decision process. This allows trade-offs between accuracy and RT to be separated from the information upon which a decision is made. For predictive inferences, this means that the degree to which a reader encodes the information that the actress will die can be measured independently of the speed/accuracy criteria that the reader adopts. With the criteria separated away from the information that readers know, we can address directly the degree to which the inference that the actress died is encoded. In the following section, we show how the model does this. The power of this approach has been demonstrated for elderly adults. Generally, prior to application of the diffusion model, it had been assumed that all, or almost all, cognitive processes slow with age and so less information can be brought to bear on decisions. However, application of the model showed this is not the case for many paradigms (Ratcliff, Thapar, & McKoon, 2004; Ratcliff et al., 2010, 2011). In these paradigms, elderly adults are slower than young adults because they set their speed/accuracy settings to value accuracy more than young adults do.

The paradigm we used in the study reported here was single-word recognition memory, a task that is frequently used in cognitive psychology and one that was used by McKoon and Ratcliff (1986). On each of a series of read-and-test lists, participants are asked to read several sentences, unrelated in content to each other, and then they are tested on a list of single words. For each word, they are asked to decide if it had or had not appeared in any of the sentences just read, as quickly and accurately as possible. A test word that is related to a just-read sentence is separated from that sentence by the other sentences to be read and other test words, so the memory being tested is long-term memory for the information that was read in a sentence. For college students, RTs on this task are in the range of 600–800 ms and accuracy is in the range of 70%–90%.

Among the sentences used in this study were predictive inference sentences, and for these sentences, the test words of interest were words that denoted the events predicted by the sentences. For the actress sentence, the target test word was "dead." Another example is the sentence "The mover bent his knees, put his arms around the box, and took a deep breath," for which the target test word was "lift." There were two versions of each sentence: One predicted the target event, and the other was a control sentence, a sentence that used as many of the same words as the predicting sentence as possible without predicting the target event ("The director fell upon the cameraman, demanding that he get close-ups of the actress on the 14th story"; "The mover laid down the box, rested his knees and arms, and caught his breath"). Using the same words in the predicting and control sentences was intended to equate semantic associations between the individual words and the target words.

The target test words did not appear explicitly in the sentences that were read, and so the correct response was "no." The logic was that, to the extent that a participant had encoded the appropriate inference when a predicting sentence was read, it would be difficult to respond "no" to the target test word—the meaning of the target test word would, at least partially, match the encoded meaning of the sentence and lead to a tendency for participants to respond "yes," incorrectly. The match may be only partial because, as McKoon and Ratcliff (1986, 1992) showed, the inference may be only minimal, for example, "something bad happened" rather than "death."

We used a memory task in order to assess whether readers had encoded the predictive inferences during reading of the sentences. There are other paradigms with which to investigate the generation of inferences, but they do not give evidence that the inferred information is encoded into memory (e.g., McKoon & Ratcliff, 1986, 2015). In one frequently used task, test words are presented immediately after a sentence is read (e.g., "dead" presented immediately after an actress sentence). In some studies, the words

are presented for recognition ("Was this word in the sentence you just read?"); in others, they are presented for lexical decision; and in others, they are presented for naming aloud. Online tests like these cannot show whether an inference was encoded during reading of a sentence; instead they may show only information that was available from the interaction of the test word and the sentence (e.g., Forster, 1981; McKoon & Ratcliff, 1994; McKoon, Ratcliff, & Ward, 1994; Ratcliff & McKoon, 1988); without the test word coming immediately after the sentence, no inference would be encoded. The same problem exists when sentence reading times are measured. If two sentences are related to each other and they are read one right after the other, the reading time for the second sentence may reflect only information that was generated when the second sentence was read, not information that was generated during reading of the first sentence. For these reasons, and because our goal was to determine whether predictive inferences are actually encoded during reading, we used the single-word recognition memory task. We note that responses to test words in this task reflect fast, automatic, passive memory processes, not slow and strategic ones. Cook and O'Brien (2015), McKoon and Ratcliff (1989), and Posner (1978) gave a more complete discussion of methods of testing for inferences.

We should point out here that our investigation is different in an important way from other research often conducted with low-literacy adults in which the aim is to predict overall performance on tests of comprehension rather than to understand what particular aspects of comprehension give difficulties. There are many individual processes involved in comprehending a text, ranging from understanding the contextually relevant meanings of words to understanding the structure of a text as a whole. Our aim is to investigate isolated processes like these in contrast to measures that collapse over them, such as the number of multiple-choice questions answered correctly about a text just read or the number of blank words in a text that can be correctly filled in.

The diffusion model

We focus on the diffusion model because it allows the knowledge an individual has encoded from a text to be separated from the individual's speed/accuracy choices. To our knowledge, such a model-based separation is new to research with low-literacy adults.

The dependent variables in our experiment were RTs and accuracy for the test words. However, participants, or groups of participants, cannot be directly compared in terms of these variables. One reason, just given, is that there can be differences among individuals in speed/accuracy trade-off settings. Another reason is that individuals may have different baseline levels of performance. In a single-word memory test, low-literacy individuals might have a 50-ms difference between conditions on a baseline of 800 ms and college students a 25-ms difference on a baseline of 700 ms; low-literacy individuals might have a 2% difference in accuracy on a baseline of 95% and college students a 5% difference on a baseline of 90%. Again, interpretation of results like these can be done only if individuals' speed/accuracy settings are known; in particular, it cannot be determined whether the information on which low-literacy adults base their responses reflects encoded inferences.

The diffusion model (Ratcliff, 1978; Ratcliff & McKoon, 2008) directly measures the strength with which information is encoded in memory (e.g., the strength with which death is encoded after an actress sentence) because it separates strength from speed/accuracy criteria (and encoding and response execution processes). The model applies to decisions that are made quickly, in under 1 or 2 s, as responses in single-word tests of recognition memory are for both low-literacy adults and college students.

For single-word test items, a test word must be encoded and must be transformed into a representation of the information on which a decision is based. In the case here, it is the strength of the memory representation of the word (not, e.g., its font size or font color). Then, given that information, a decision is made and a response executed. For the application here, it is the decision process that is of interest, so encoding and response execution processes are combined into one parameter of the model, which we call *nondecision* time.

The model as it applies to single-word recognition is illustrated in Figure 1. Total RT is the sum of the time taken by the nondecision processes (*Ter*) and the time taken to make a decision. The top panel

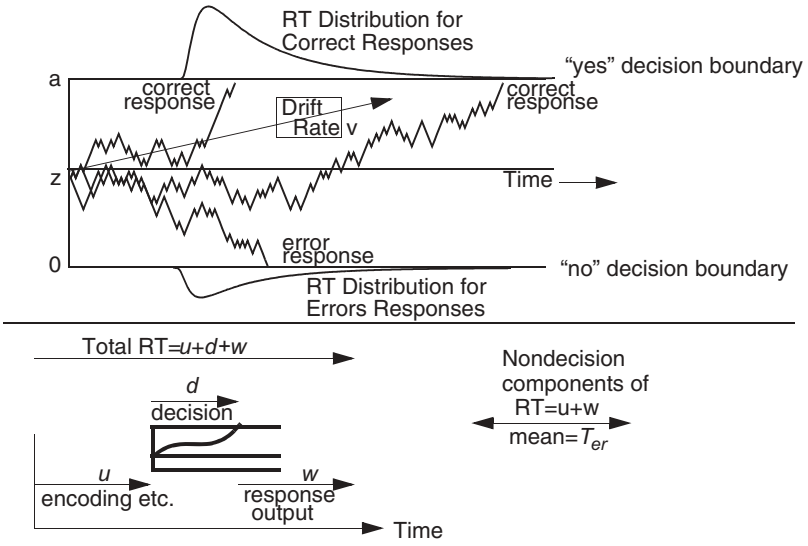

Figure 1. An illustration of the diffusion process. *Note*. The top panel shows three simulated paths with mean drift rate *v*, starting point *z*, and boundary separation *a*. One process hits the top boundary quickly, another hits it later, and another hits the bottom boundary in error. The top panel also shows how the model predicts the right-skewed shapes of response time (RT) distributions: Most processes hit the boundary quickly, but some hit later. Drift rate is normally distributed across trials with standard deviation η, starting point is uniformly distributed with range s_z, and nondecision time is uniformly distributed with range s_t.

illustrates the decision process. Information from memory about a test word is accumulated over time from a starting point (z) to one or the other of two criterial amounts, or boundaries, one for each choice. The stronger the information from memory, the higher the rate at which it is accumulated. The rate of accumulation is called drift rate, *v*. Test words that differ in difficulty differ in drift rates. For example, in the actress sentences, responses might be easier for "actress" as a test word than for "story" as a test word because the actress is the topic of a clause. Drift rates for test words that did appear in a sentence have positive values, moving, on average, from the starting point toward the upper boundary (*a*), and test words that did not appear in any sentence have negative values, moving, on average, from the starting point toward the lower boundary (*0*). A response is executed when the amount of accumulated information reaches a boundary, *0* or *a*.

The process of accumulating information is noisy. Three instances are shown in the figure. They have exactly the same mean drift rate, but noise means that they approach the boundaries at different rates and sometimes that they reach the wrong boundary. The noise leads to the distributions of RTs shown in the figure, with more shorter than longer RTs (which is the distribution shape almost always found empirically). We call this noise *within-trial* variability.

The stimuli of a particular condition in an experiment are assumed to all have the same mean value of *v*. For example, for an experiment with paragraphs as the texts to be read, the test words in one condition might be words that were highly topical in a paragraph and the test words in the other condition might be words that were only details; the value of *v* would be larger for the first condition than the second. However, the value of *v* for a particular condition is assumed to vary around its mean from trial to trial. This assumption comes from the notion that an individual cannot hold the mean value of *v* exactly the same from one trial of a condition to the next (Ratcliff, 1978). For the same reason, it is assumed that there is trial-to-trial variability in the starting point (Laming, 1968) and in the nondecision time (Ratcliff & Tuerlinckx, 2002). We call the variability from trial to trial *across-trial* variability. Across-trial variability in drift rate is assumed to be normally distributed with standard deviation η, across-trial variability in the nondecision component is assumed to be uniformly distributed with range st, and across-trial variability in the starting point is assumed to be uniformly distributed with range s_z. (These assumptions are not crucial because the model is robust to the forms of the distributions.)

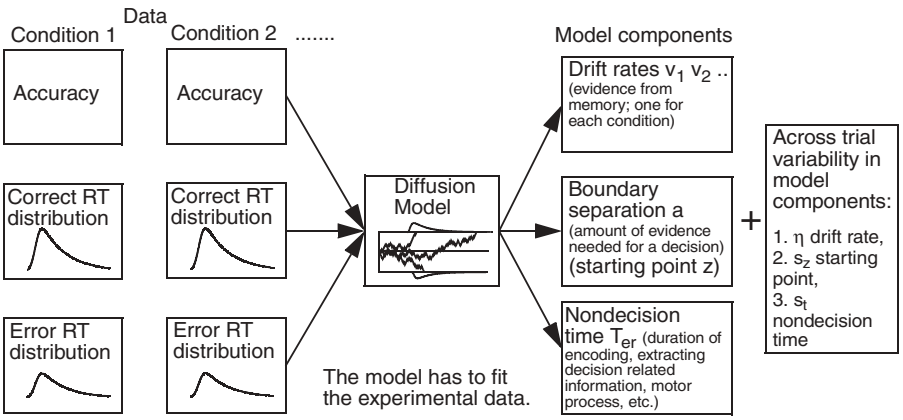

Figure 2. An illustration of the mapping from response time (RT) distributions and accuracy to drift rates, boundary settings, and nondecision time.

The diffusion model is validated only if it explains all aspects of data: accuracy, mean correct RTs, mean error RTs, the shapes and locations of RT distributions, and the relative speeds of correct and error responses. As shown in Figure 2, the model maps these data to the underlying components of processing: drift rate (*v*), starting point (*z*), boundary separation (*a*), and nondecision time (*Ter*). In the figure, the data, on the left, are mapped through the model to give the values of the components, on the right.

The reason in the model that an individual's accuracy cannot be predicted from his or her speed, or vice versa, is that drift rates and boundary settings are independent of each other. For example, for a given value of accuracy, an individual might set his or her boundaries close together and so respond quickly, or he or she might set them farther apart and so respond more slowly.

We stress that the model is tightly constrained. The first and most powerful constraint comes from the requirement that the model fit the right-skewed shape of RT distributions that is almost always obtained (Ratcliff, 1978, 2002; Ratcliff & McKoon, 2008; Ratcliff, Van Zandt, & McKoon, 1999). Second, across experimental conditions that vary in difficulty (and are randomly intermixed at test), changes in accuracy, RT distributions, and the relative speeds of correct and error responses must all be captured by changes in only one parameter of the model, drift rate. The boundaries cannot be adjusted as a function of difficulty because it would be necessary for the system to know which level of difficulty was being tested before the drift rate could be determined. Third, across experimental conditions that vary in speed/accuracy criteria (e.g., speed vs. accuracy instructions), all the changes in accuracy, RT distributions, and the relative speeds of correct and error responses must be captured by changes in the settings of the response boundaries (and, empirically, there are also sometimes small changes in nondecision time).

Important to note, the model is identifiable and falsifiable. Although it is relatively easy for the 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 must meet the three constraints just mentioned. Also, Ratcliff (2002) made up several sets of fake but quite plausible data and showed that the diffusion model failed (dramatically) to fit them. (Note that in most comparisons made so far, e.g., Ratcliff, Thapar, Smith, & McKoon, 2005, conclusions drawn from models other than the diffusion model, e.g., Ratcliff, Thapar, & McKoon, 2007; Usher & McClelland, 2001—see also Donkin, Brown, Heathcote, & Wagenmakers, 2011—have been the same as those from the diffusion model.) In the cognitive psychology literature, there has recently been a discussion about the identifiability of the diffusion model (and sequential sampling models in general). Jones and Dzhafarov (2014) implied that all models in which evidence is accumulated over time are unfalsifiable in that they can account for any pattern of data. But their argument was misleading—it was limited to models that are deterministic or very nearly deterministic (Smith, Ratcliff, & McKoon, 2014). If a model is stochastic, that is, if it assumes that there is noise in the decision process and that the noise

gives rise to correct and error responses and responses with varying RTs (i.e., RT distributions), as the diffusion model does, then the model is falsifiable.

The diffusion model allows performance to be compared among individuals because the differences among individuals in drift rates, boundary settings, and nondecision time are typically 3 to 5 times larger than the sampling variability in the model parameters (with only a single session of data; Ratcliff et al., 2010; Ratcliff et al., 2015). Ratcliff and Tuerlinckx (2002; Ratcliff & Childers, 2015) provided analyses and tables that allow power and effect sizes to be computed. Explaining data at the level of individuals is essential for comparisons among them and for comparisons of one group of individuals to another.

The diffusion model also increases power when there are relatively small numbers of observations in the critical conditions of an experiment. Filler conditions with many more observations allow accurate estimation of all the parameters of the model other than drift rates, and these simultaneously constrain drift rates for the critical conditions giving an increase in power (White, Ratcliff, Vasey, & McKoon, 2010). This methodology was applied to the data from the experiment described here.

As just mentioned, the differences between interpretations of data provided by the model and interpretations generated from dependent variables directly are illustrated by studies with older adults, ages 65 to 90, and college students (Ratcliff et al., 2010, 2011; Ratcliff et al., 2015). For the single-word memory task, there are large increases in RTs with age and only small changes in accuracy or no changes in accuracy at all. The RT data suggest large decrements in memory with age, but the accuracy data suggest only small decrements or none. The diffusion model reconciles these seemingly inconsistent results by mapping the two dependent variables onto the same underlying mechanisms. In many but not all tasks, the large increases in RTs with age are due mainly to wider boundary settings and the duration of the nondecision processes. The small or nonexistent deficits in accuracy are due to small or nonexistent decreases in drift rates. These findings (Ratcliff et al., 2004, 2010, 2011) show that older adults' slow performance in the single-word memory task is due only negligibly to poor memory, counter the usual claim that older adults have significantly worse memory than young adults.

The single-word memory task and diffusion model analyses of the data from it are fundamentally different from some of the psychometric tasks and analyses often used to investigate reading comprehension for low-literacy adults. Often, abilities such as IQ, working memory, and spelling are tested to determine which combinations of them predict reading comprehension, as measured by some test of global comprehension (e.g., answering multiple-choice questions about a just-read passage). With this approach, individual difference variables like IQ and standardized test scores are entered into analyses. There is no way to know if the score from a particular test estimates exactly what it is meant to measure (except by evaluating validity with respect to other tests or outcome measures). Often, the questions addressed with these methods have been concerned with theoretical constructs that have been hypothesized to relate to reading, for example, rapid automatic naming, phonemic decoding, auditory working memory, word reading, reading fluency, vocabulary, and comprehension of spoken language were hypothesized by Mellard, Fall, and Woods (2010); decoding, word recognition, spelling, fluency, and comprehension by MacArthur, Konold, Glutting, and Alamprese (2010); and phonological processing, word reading, spelling, vocabulary, processing speed (fluency), and cognitive ability by Mellard, Woods, Desa, and Vuyk (2013). For each construct, tasks are chosen to measure it, and the scores on the tasks are entered into, for example, confirmatory factor analysis, path analysis, or principal axis factoring. In most of the tasks, only accuracy is measured; there is no measure of the speed of responses for individual test items. As said earlier, without a measure of speed, the effects of various abilities on task performance cannot be unequivocally interpreted.

Experiment

The materials were pairs of sentences, each made up of a predictive sentence and its control, like the actress and mover sentences. For each pair, we attempted to ensure, as much as possible, that the control sentence would not be associated with the predictive inference. As the data presented later indicate, we did not completely succeed in this. For the lift sentences, for example, "lift" might be

associated to the control sentence ("The mover laid down the box, rested his knees and arms, and caught his breath") even though it is not a predicted event. Nevertheless, the data from the experiment show that the inferences generated from the predictive sentences were stronger than the associations between target words and control sentences.

There were two groups of participants in the experiment. For one, the participants were students in Adult Basic Literacy Education (ABLE) classes in the Columbus, Ohio, area, which are free and open to anyone. For the other, the participants were college students. We included the college students to check that our materials, modest revisions of those used by McKoon and Ratcliff (1986), allowed replication of McKoon and Ratcliff's findings, showing that the materials give evidence of predictive-inference encoding when it does, in fact, occur. The two groups could not, of course, be matched on such demographic and cognitive variables as IQ, age, social-economic status, or performance on standardized comprehension tests. To preview, we found that the low-literacy participants showed encoding for the inferences, despite their poor performance on a standardized comprehension test.

Method

Materials

There were 48 pairs of predicting and control sentences. The test words for these sentences were the person engaged in the predicted event (e.g., "actress," "mover"), a word representing the predicted event (e.g., "dead," "lift"), and three other content words. The same words were tested for the control sentences as the predicting sentences. There were also 96 filler sentences, five test words from each, and a pool of 2,149 words that did not appear in any sentence from which test words were selected randomly for each participant.

We checked that none of the sentences used content words that were unlikely to be familiar to ABLE students. To do this, all the content words were tested in a lexical decision task with ABLE students who did not participate in the experiment described here. For each of a series of strings of letters, they were asked to decide as quickly and accurately as possible whether the string was or was not an English word. Words for which accuracy was less than 80% were not used in this experiment. Although lexical decision responses may not reflect the full meanings of words and they may not directly assess the "sluggishness" of word-level skills (Perfetti & Stafura, 2014, 2015; Perfetti, Yang, & Schmalhofer, 2008), they do require at least some level of familiarity with the words.

Procedure

Stimuli were displayed on a PC screen, and responses were collected from the PC's keyboard. The experiment began with 30 lexical decision test items, used for practice with the PC keyboard. Then there were 48 blocks of sentences to be read with their test words (preceded by one practice block, which contained only filler sentences). For each of the blocks, there were three sentences, two filler sentences, and one experimental sentence (i.e., the predicting or control sentence from one of the pairs). The experimental sentence was always the second of the three. Participants began each block by pressing the space bar on the keyboard. Then the sentences were displayed one at a time for 4,500 ms, 7,000 ms, or 10,000 ms, depending on whether they were one, two, or three lines long as displayed on the PC screen. There was a 500-ms blank screen between sentences.

The three sentences were followed by 27 test words, 14 from the three sentences just read and 13 that did not appear in any of the sentences in the experiment (one of which was a target test word, e.g., "dead," "lift"). For each word, participants were asked to respond "yes" or "no" as quickly and accurately as possible according to whether the word had appeared in any of the just-read sentences, using the ?/ key on the PC keyboard for "yes" responses and the Z key for "no" responses. Each test word was displayed until the participant made a response, and then the screen was cleared. If the response was correct, the next test word appeared in 300 ms. If the response was not correct, the word ERROR appeared for 900 ms

and then the screen was cleared for 300 ms. The test words were presented in random order, except that the target test word for a predicting or control sentence was immediately preceded by a word from its sentence (e.g., "actress," "mover").

The 48 pairs of experimental sentences were counterbalanced across sentences and participants such that half of the sentences for each participant and half of the participants for each sentence were in the predicting condition and half in the control condition.

Participants

Twenty-eight undergraduates from Ohio State University took part in the experiment for credit in an introductory psychology class, and 28 students participated from ABLE classes in the Columbus area.

The ABLE students varied from 28.5 years old to 80.2 years old (M = 48.6, SD = 11.9). Their reading levels were measured by the TABE test, a widely used test of adults' reading ability that is used in the ABLE classes to determine reading grade levels. 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. Scores on this test are translated into reading grade levels. For the ABLE students in our study, the students' grade levels varied from 2.5 to 12.9 (M = 6.6, SD = 2.8). The ABLE students in the experiment reported here had all participated in a previous lexical decision study that did not include any of the test words of interest in this experiment. (Note that the TABE test seems quite reliable in the context here: McKoon & Ratcliff, 2016, showed that scores on the TABE were correlated with drift rates from a lexical decision task 0.48.)

Results

Mean RTs and accuracy values are shown in Table 1. RTs longer than 3,000 ms and shorter than 300 ms were eliminated from the analyses. This was 2.2% of the data for the ABLE students, with 1% being fast guesses from three participants. For the college students, 7.4% of the data were eliminated, with 5.8% of this coming from fast guesses from four participants. There were seven findings of interest:

(1) Accuracy showed the anticipated difference between the predicting and control conditions for both groups: It was more difficult to respond "no" correctly to target test words in the predicting than the control condition.
(2) The difference in accuracy between the predicting and control conditions was much larger for the college students (.21) than the ABLE students (.08).
(3) Averaging the probability of "yes" responses for the predicting, control, filler positive, and filler negative conditions, the ABLE students were more likely to respond "yes" (.64) than the college students (.48).
(4) The probability of a "yes" response to target words in the control condition was higher than the probability of a "yes" response to other test words that had not appeared in the sentences. We attribute this to overlap in meaning between the control sentences and the target words (e.g., overlap between "laid down box," "arms," knees," and "lift"), that is, we did not succeed as much as we would have liked in keeping the control sentences completely unrelated to the target test words.
(5) Incorrect "yes" responses for target test words were faster in the predicting than the control condition for both groups.
(6) This difference in RTs was 32 ms on a baseline of 783 ms for the college students and 43 ms on a baseline of 895 ms for the ABLE students.
(7) The ABLE students' responses were slower than the college students' for all test words.

Table 1. Accuracy and Mean RTs (ms)

Statistic	Participant group	"Yes" responses, predicting condition	"Yes" responses, control condition	"Yes" responses, filler positives	"No" responses, filler negatives
Response proportion	College	0.60	0.39	0.81	0.87
	ABLE	0.76	0.68	0.80	0.67
Mean RT forcorrect responses	College	751	783	754	794
	ABLE	852	895	862	999

The main analyses of the data were those using the diffusion model. However, for completeness, we give the results of simple analyses of variance (ANOVAs). We conducted a two-factor ANOVA of the accuracy data with group of participants as one factor and condition as the other, where the conditions were target test words in the predicting condition, target test words in the control condition, and filler positive test words. The tendency for the ABLE students to respond "yes" with higher probability than the college students was significant, $F(1, 54) = 44.1$, $p < .05$, is used throughout this article ($\eta^2 = 0.15$, MSE = 0.916), as was the main effect of condition, $F(2, 108) = 80.8$ ($\eta^2 = 0.33$, MSE = 0.975), but these were qualified by an interaction, $F(2, 108) = 28.0$ ($\eta^2 = 0.11$, MSE = 0.337).

For the interaction, post hoc tests showed that the probability of responding "yes" to a target word was larger for the predicting than the control condition, $F(1, 54) = 49.0$, but the difference between the two conditions (finding 2 above) was much larger for the college students than the ABLE students, $F(1, 54) = 110.0$. This is a scaling problem: For these conditions, the college students were operating at a baseline of about .50 probability of "yes" responses, whereas the ABLE students' baseline was about .72. This means that the predicting-control difference cannot be directly compared between the two groups. A difference of .21 on a baseline of .50 cannot be directly compared to a difference of .08 on a baseline of .72. Taking *z* scores of these probabilities to correct for baselines reduces the differences to 0.53 and 0.24, but there is still a larger difference for college students than ABLE students. However, in the following analyses for which accuracy and RTs are jointly considered and translated into drift rates, the difference between the two groups of students disappears.

The ABLE students' tendency to respond "yes" with greater probability than the college students is especially apparent when the probabilities for the target words in the control condition are compared to those for the negative filler words. The college students' difference was .26, whereas the ABLE students' difference was .01; the difference between these two probabilities was significant, $F(1, 54) = 156.3$. However, as the diffusion model analyses shows, this did not prevent the ABLE students from encoding predictive inferences.

We conducted the same ANOVA for RTs. The ABLE students were significantly slower than the college students, $F(1, 54) = 8.8$ ($\eta^2 = 0.12$, MSE = 527789); there was a significant difference across the conditions, $F(2, 108) = 4.1$ ($\eta^2 = 0.01$, MSE = 24665); and there was no significant interaction, $F(2, 108) = 0.6$ ($\eta^2 = 0.002$). This lack of interaction is not meaningful because of the baseline differences in the RTs of the two groups (finding 6 above). We leave further interpretation of these effects to the diffusion model analyses.

Diffusion model analyses

As discussed earlier, the model needed to account for accuracy, the shapes and locations of the distributions of RTs for correct responses and for errors, and the relative speeds of correct and error responses, and it needed to do this simultaneously for the two groups of participants and all the conditions in the experiment.

To fit the model to the data, the RT distributions were represented by five quantiles, the .1, .3, .5 (the median), .7, and .9 quantiles. The model was fit with a chi-square minimization method that is fully described in Ratcliff and Tuerlinckx (2002). The correct and error RT distributions were fit for all the

conditions simultaneously, and they were weighted by the number of observations (because the chi-square method uses frequencies). The model was fit to the data for each participant individually.

We divided the data into eight categories to increase constraints on the model and to provide parallel analyses to those in a similar study with elderly adults (McKoon & Ratcliff, 2013). Together with the numbers of observations per participant, the categories were as follows: target test words in the predicting condition (24) and the control condition (24), words from filler sentences (480), words that did not appear in any sentence (excluding target test words; 576), words that immediately preceded the target words in the predicting condition (24) and the control condition (24), and filler words from the predicting sentences (72) and control sentences (72). For statistical analyses, we chose the positive filler category to compare with the predicting and control categories because it had the largest number of observations for test words for which the correct response was "yes" and so was most reliable among those categories.

The model fit the data well, as shown by the chi-square values in Table 2. The chi-square test is a very conservative test, so even when chi-square values are lower than twice the critical value, the fit of the model to data is good (see Ratcliff et al., 2004, for a discussion of model fitting and chi-square values). In fact, of the chi-square values for individual participants, only two in each group had chi-square values greater than twice the critical value. For comparing the chi-square values to those obtained by McKoon and Ratcliff (2013), note that the number of observations per participant in the study reported here was about 50% larger, which means that differences in chi-square due to small, systematic misfits are magnified. The best-fitting values for all the model's parameters except drift rates are given in Table 2 and those for drift rates in Table 3.

The quality of the fit of the model to the data is also shown in the plots in the appendix, where the x-axis is the data and the y-axis is the model's values. The first plot shows the data and fits for

Table 2. Diffuson model parameters and chi-square values.

Participant group and statistic	a	T_{er} (ms)	η	s_z	p_o	s_t (ms)	z	χ^2
ABLE mean	0.156	0.577	0.161	0.049	0.002	0.256	0.097	116.6
College mean	0.136	0.501	0.170	0.042	0.004	0.215	0.075	119.1
ABLE SD	0.030	0.076	0.071	0.042	0.002	0.097	0.023	49.5
College SD	0.024	0.049	0.068	0.026	0.005	0.098	0.016	41.7

Note. a = boundary separation, z = starting point, T_{er} = nondecision time, η = standard deviation in drift across trials, s_z=range of the distribution of starting point (z)across trials, s_t = range of the distribution of nondecision times across trials. Degrees of freedom for the χ^2 were computed as follows: for the 5 quantile RTs for correct responses and the 5 quantile RTs for error responses, there were 6 bins (2 outside the .1 and .9 quantiles and 4 between the pairs of quantiles). This gives 12 degrees of freedom for each category of test words minus 1 (because the 12 numbers must sum to 1). With 11 degrees of freedom for each of the eight categories, there was a total of 88 degrees of freedom. There were 15 model parameters (shown in Table 2 and 3) and so the number of degrees of freedom was 88-15=73, for which the critical value is 93.9. There were no significant differences between the college and ABLE students in the across-trial variability parameters. In almost all RT studies, some proportion of responses, p_o, are spurious "contaminants" (that might arise from inattention for example). These are modeled as a mixture of pure decision processes and pure decision processes to which a delay has been added. (Ratcliff & Tuerlinckx, 2002). The probability of contaminants (p_o) was less than .5% and so they are not considered further.

Table 3. Diffusion model drift rates.

					Word immediately preceding target			
Participant Group and Statistic	Target in predicting condition	Target in control condition	Words from filler sentences	Negative words	Predicting condition	Control condition	Filler from predicting sentence	Filler from control Sentence
ABLE mean	0.147	0.072	0.128	-0.143	0.211	0.195	0.100	0.109
College mean	0.038	-0.062	0.178	-0.249	0.250	0.274	0.130	0.135
ABLE SD	0.149	0.164	0.055	0.122	0.140	0.122	0.063	0.083
College SD	0.070	0.073	0.052	0.095	0.124	0.112	0.043	0.047

proportions of responses for correct responses. There are few deviations greater than 5% for the conditions with more than 24 observations, which shows good match between theory and data. For RTs, the next three plots show the 0.1 RT quantile (representing the leading edge of the RT distributions), the 0.5 quantile (representing the median), and the 0.9 quantile (representing the tail). The x's come from conditions with smaller numbers of observations (more than seven and less than 25) and the o's from conditions with larger numbers of observations (25 or more). There are more than 200 observations per plot (28 participants with eight conditions per participant), so even though a few deviations appear large (more for the low numbers of observations, i.e., more for the x's than the o's), overall the fits are good. For the 0.1 quantile RTs from conditions with large numbers of observations, there were five deviations greater than 50 ms for the ABLE students and somewhat more for the college students. For the median RTs there were few serious deviations, and for the 0.9 quantiles there was somewhat more variability.

Implicit in the good quality of the fit of the model to the data is that it separated those elements of responding that mainly determine RTs from the information that is available from memory. The differences in RTs between the ABLE and college students were mainly due to differences in their boundary settings and nondecision times (Table 2). The distance between the boundaries (a) was larger for the ABLE participants, indicating that they were more conservative, $t(51.2) = 2.8$, $\eta^2 = 0.13$, confidence interval on the difference 0.006, 0.135, and their nondecision processes were slower, $t(46.3) = 4.4$, $\eta^2 = 0.26$, confidence interval on the difference 0.041, 0.109.

With boundary settings and nondecision times abstracted away, the information in memory that was encoded from sentences can be directly compared between the ABLE and college students (Table 3). For all the conditions except the target test words, the ABLE students had lower absolute values of drift rates than the college students. Also, the difference between drift rates for words from filler sentences and negative words was larger for the college students—.427, .178, .249—than the ABLE students, .271, .128, .143. These results show the college students' better memory for information from the sentences.

To examine drift rates for the target test words, we began by conducting an ANOVA on drift rates with two factors: the two groups of participants and three of the conditions—target test words in the predicting and control conditions and filler positive test words. The difference between the groups was significant, $F(1, 54) = 11.4$ ($\eta^2 = 0.06$, MSE = 0.171), with the ABLE students more likely to respond "yes"; the difference among the conditions was significant, $F(2, 108) = 35.2$ ($\eta^2 = 0.22$, MSE = 0.310), and the interaction was significant, $F(1, 108) = 15.8$ ($\eta^2 = 0.09$, MSE = 0.139).

The first and most important result from this analysis was that there was no significant difference between the *relative* degree to which the two groups encoded target test words. The difference between the drift rates for the predicting and control conditions was .075 for the ABLE students and .100 for the college students, a difference that was not significant, $F(1, 54) = 1.8$. Although we cannot conclude that there is no difference between the two groups, we can conclude that the difference must be very small.

This gives a conclusion that is quite different from the one that would be drawn from the accuracy data. In accuracy, the difference between the predicting and control conditions was much less for the ABLE students than the college students, 0.08 for the ABLE students and 0.21 for the college students. The model is fit jointly to all conditions and all the conditions determine the model parameters that are common to all conditions (boundary separation, nondecision time, variability conditions). This provides more power than just one of the dependent variables (see White et al., 2010).

The second result was that the ABLE students' responses, compared to the college students', tended to be based on the general compatibility of a test word with memory for its sentence (e.g., "lift" associated with the words in the control sentence). The ABLE students' drift rates did not significantly differentiate between, on one hand, target test words in the predicting and control conditions, words that were not presented explicitly in the text ($M = .109$), that is, (.147 + .072)/2), and, on the other hand, test words from filler sentences ($M = .128$), $F(1, 54) = 1.0$. In contrast, the college students' did differentiate between the

target test words in the predicting and control conditions and test words from filler sentences (M for predicting and control conditions = –.012), that is, (.038-.062)/2 (M for fillers = .178), $F(1, 54) = 53.3$. This difference between the two groups seems to come about because the ABLE students had more difficulty overall than the college students and were matching meaning rather than veridical information; nevertheless, the difference they showed between predicted and control test words was significant.

Discussion

Reading comprehension is composed of many processes and many interactions among them. The particular kind of inference we investigated, predictive "what happens next" inferences, has not been investigated previously with low-literacy adults, despite its pivotal contribution to the development of reading comprehension theories.

Our finding that the degree to which low-literacy adults encoded predictive inferences was not significantly different from college students' contrasts with the long list of studies that have found that low-literacy adults encode inferences to a lesser degree. There are (at least) two possible reasons for this. One is simply that we investigated a different type of inference, and the other is that we used the diffusion model to separate the speed/accuracy trade-offs that individuals adopt from the information they encode during reading. Whether and which other types of inferences are encoded by low-literacy adults is a question for future research.

The range of the reading grade levels for the ABLE students in our experiment was wide, 2.5 to 12.9 (M = 6.6). This suggests another avenue for future research, measuring reading skills at the level of individual students and examining correlations between scores on standardized reading comprehension tests, the information the students understand from texts as measured by drift rates in the diffusion model, and individual differences in, for example, IQ or age. We are currently engaged in these sorts of analyses with a pool of 100 ABLE students.

Perhaps the most important of our findings is that the diffusion model was used successfully to interpret data for low-literacy adults. As previously emphasized, an individual's overall speed and accuracy are largely independent; his or her accuracy cannot be predicted from his or her speed, and his or her speed cannot be predicted from his or her accuracy. In addition, the model can be used to evaluate the performance of single participants, as it successfully did in the study reported here. One import of this is that values of drift rates, boundary settings, and nondecision processes can be correlated with scores on tests of other kinds of inferences, individual-difference measures like IQ, age, and social/economic status, and scores on tests of comprehension that are more global such as the TABE and question-answering paradigms.

The divergence of RTs and accuracy for the interpretation of data was apparent in the data reported here. RTs and accuracy were not separately interpretable; the ABLE students' RTs were 100–200 ms longer than the college students', and their accuracy was about 10% lower. The conclusion about predictive inferences from the accuracy data alone would have been that ABLE students do not encode them to the extent that college students do; the difference in accuracy between the predicting and control conditions was .21 for the college students but only .08 for the ABLE students. From the RT data alone, the results would have been equivocal; the difference between the predicting and control conditions was 32 ms for the college students and 43 ms for the ABLE students, but they were on quite different baselines: 783 ms and 895 ms, respectively.

These results pose a serious challenge to much of the work that has been done to examine inference processing for low-literacy adults. It may be that new studies are needed that examine both RTs and accuracy as dependent measures. It may be that differences in inference processing between adults with low literacy skills and adults with good skills are considerably less than has been thought previously. If so, new hypotheses about reading difficulties may emerge.

One way to describe our results is that the ABLE students knew more than their accuracy scores indicated. To carry this further, suppose that a correct response for some item on some global comprehension test (e.g., a multiple-choice test) requires knowing about a predicted event like death.

ABLE students would, following our results, be less accurate than college students, and in the absence of RTs, it would be concluded that they knew less than the college students about the predictive inferences that could be encoded from a text.

The possibility that individuals know more than their accuracy indicates also compromises correlations that might be obtained between accuracy and individual difference variables such as IQ, vocabulary, spelling, and the like, and it compromises comparisons from one standardized test to another. This situation is made even more difficult when standardized tests that are proprietary are used to measure comprehension, and so individual items on the test cannot be evaluated with respect to the particular skills they address. Adding further difficulty, it is clear that scores on one standardized test do not necessarily lead to the same conclusions as scores on another (Betjemann, Keenan, Olson, & DeFries, 2011; Hua & Keenan, 2014; Keenan & Betjemann, 2006).

In sum, we hope that the outlook on comprehension that we have adopted for this study, the focus on a particular kind of information necessary for full comprehension and a model-based approach to interpreting data, will be useful in identifying comprehension difficulties in a way that can lead to successful teaching methods for low-literacy adults.

Funding

Funding was provided by the Institute of Educational Sciences (R305A120189).

References

Baer, J., Kutner, M., Sabatini, J., & White, S. (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*. National Center for Education Statistics. Retrieved from http://eric.ed.gov/?id=ED505187

Barnes, M. A., Ahmed, Y., Barth, A., & Francis, D. J. (2015). The relation of knowledge-text integration processes and reading comprehension in 7th- to 12th-grade students. *Scientific Studies of Reading*, *19*(4), 253–272. doi:10.1080/10888438.2015.1022650

Betjemann, R. S., Keenan, J. M., Olson, R. K., & DeFries, J. C. (2011). Choice of reading comprehension test influences the outcomes of genetic analyses. *Scientific Studies of Reading*, *15*(4), 363–382. doi:10.1080/10888438.2010.493965

Bowyer-Crane, C., & Snowling, M. J. (2005). Assessing children's inference generation: What do tests of reading comprehension measure? *The British Journal of Educational Psychology*, *75*(Pt. 2), 189–201. doi:10.1348/000709904X22674

Cain, K., & Oakhill, J. V. (1999). Inference making ability and its relation to comprehension failure in young children. *Reading and Writing*, *11*(5–6), 489–503. doi:10.1023/A:1008084120205

Cain, K., & Oakhill, J. (2006). Profiles of children with specific reading comprehension difficulties. *British Journal of Educational Psychology*, *76*(4), 683–696. doi:10.1348/000709905X67610

Cain, K., Oakhill, J. V., Barnes, M. A., & Bryant, P. E. (2001). Comprehension skill, inference-making ability, and their relation to knowledge. *Memory & Cognition*, *29*(6), 850–859. doi:10.3758/BF03196414

Cain, K., Oakhill, J., & Lemmon, K. (2004). Individual differences in the inference of word meanings from context: The influence of reading comprehension, vocabulary knowledge, and memory capacity. *Journal of Educational Psychology*, *964*, 671–681. doi:10.1037/0022-0663.96.4.671

Casteel, M. A. (2007). Contextual support and predictive inferences: What do readers generate and keep available for use? *Discourse Processes*, *44*(1), 51–72. doi:10.1080/01638530701285572

Cook, A. E., Limber, J. E., & O'Brien, E. J. (2001). Situation-based context and the availability of predictive inferences. *Journal of Memory and Language*, *44*(2), 220–234. doi:10.1006/jmla.2000.2744

Cook, A. E., & O'Brien, E. J. (2015). Passive activation and instantiation of inferences during reading. In E. J. O'Brien, A. E. Cook, & R. F. Lorch (Eds.), *Inferences during reading* (pp. 19–41). Cambridge, UK: Cambridge University Press.

Cromley, J. G., & Azevedo, R. (2007). Testing and refining the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology*, *99*(2), 311–325. doi:10.1037/0022-0663.99.2.311

Donkin, C., Brown, S. D., Heathcote, A., & Wagenmakers, E. (2011). Diffusion versus linear ballistic accumulation: Different models but the same conclusions about psychological processes? *Psychonomic Bulletin & Review*, *18*, 61–69. doi:10.3758/s13423-010-0022-4

Forster, K. I. (1981). Priming and the effects of sentence and lexical contexts on naming time: Evidence for autonomous lexical processing. *The Quarterly Journal of Experimental Psychology Section A*, *33*(4), 465–495. doi:10.1080/14640748108400804

Garnham, A., Oakhill, J., & Johnson-Laird, P. N. (1982). Referential continuity and the coherence of discourse. *Cognition*, *11*(1), 29–46. doi:10.1016/0010-0277(82)90003-8

Gerrig, R. J., & McKoon, G. (1998). The readiness is all: The functionality of memory-based text processing. *Discourse Processes*, *26*(2–3), 67–86. doi:10.1080/01638539809545039

Gerrig, R. J., & O'Brien, E. J. (2005). The scope of memory-based processing. *Special Issue: Discourse Processes*, *39*, 225–242.

Graesser, A. C., Li, H., & Feng, S. (2015). *Constructing inferences in naturalistic reading contexts*. In E. O'Brien, A. Cook, & R. Lorch (Eds.), *Inferences during reading* (pp. 290–320). Cambridge, UK: Cambridge University Press. doi:10.1017/9781107279186.014

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

Gueraud, S., Tapiero, I., & O'Brien, E. J. (2008). Context and the activation of predictive inferences. *Psychonomic Bulletin & Review*, *15*(2), 351–356. doi:10.3758/PBR.15.2.351

Hager, C. (2012, March 21). Cat survives 19-story fall from Boston apartment building. *CBS Boston*. Retrieved from http://boston.cbslocal.com

Hua, A. N., & Keenan, J. M. (2014). The role of text memory in inferencing and in comprehension deficits. *Scientific Studies of Reading*, *18*(6), 415–431. doi:10.1080/10888438.2014.926906

Isberner, M. B., & Richter, T. (2014). Comprehension and validation: Separable stages of information processing? A case for epistemic monitoring in language comprehension. In D. Rapp & J. L. G. Braasch (Eds.), *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences* (pp. 245–276). Cambridge, MA: MIT Press.

Jones, M., & Dzhafarov, E. N. (2014). Unfalsifiability and mutual translatability of major modeling schemes for choice reaction time. *Psychological Review*, *121*(1), 1–32. doi:10.1037/a0034190

Keenan, J. M., & Betjemann, R. S. (2006). Comprehending the gray oral reading test without reading it: Why comprehension tests should not include passage-independent items. *Scientific Studies of Reading*, *10*(4), 363–380. doi:10.1207/s1532799xssr1004

Kristof, N. (2014, October 25). The American dream is leaving America. *The New York Times*. Retrieved from http://www.nytimes.com

Kutner, M., Greenberg, D., & 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: National Center for Education Statistics.

Laing, S. P., & Kamhi, A. (2003). Alternative assessment of language and literacy in culturally and linguistically diverse populations. *Language, Speech, and Hearing Services in Schools*, *34*(1), 44–56. doi:10.1044/0161-1461(2003/005)

Laming, D. R. J. (1968). *Information theory of choice-reaction times*. New York, NY: Wiley. doi:10.1002/bs.3830140408

Lassonde, K. A., & O'Brien, E. J. (2009). Contextual specificity in the activation of predictive inferences. *Discourse Processes2*, *46*, 426–438. doi:10.1080/01638530902959620

Linderholm, T., & Van den Broek, P. (2002). The effects of reading purpose and working memory capacity on the processing of expository text. *Journal of Educational Psychology*, *94*(4), 778–784. doi:10.1037/0022-0663.94.4.778

Lockhart, R. S., Craik, F. I. M., & Jacoby, L. (1976). Depths of processing, recognition, and recall: Some aspects of general memory system. In J. Brown (Ed.), *Recall and recognition*. London, UK: Wiley.

Long, D. L., & Golding, J. M. (1993). Superordinate goal inferences: Are they automatically generated during comprehension? *Discourse Processes*, *16*(1–2), 55–73. doi:10.1080/01638539309544829

Long, D. L., Oppy, B. J., & Seely, M. R. (1994). Individual differences in the time course of inferential processing. *Journal of Experimental Psychology. Learning Memory and Cognition*, *20*(6), 1456–1470. doi:10.1037//0278-7393.20.6.1456

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*(2), 108–121. doi:10.1177/0022219409359342

McFadden, R. D. (2007, December 8). Two window washers fall 47 floors. *The New York Times*. Retrieved from http://www.nytimes.com

Magliano, J. P., & Millis, K. K. (2003). Assessing reading skill with a think-aloud procedure and latent semantic analysis. *Cognition and Instruction*, *21*(3), 251–283. doi:10.1207/S1532690XCI2103_02

McKoon, G., & Ratcliff, R. (1986). Inferences about predictable events. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *12*(1), 82–91. doi:10.1037/0278-7393.12.1.82

McKoon, G., & Ratcliff, R. (1989). Inferences about contextually defined categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(6), 1134–1146. doi:10.1037/0278-7393.15.6.1134

McKoon, G., & Ratcliff, R. (1992). Inference during reading. *Psychological Review*, *99*(3), 440–466. doi:10.1037/0033-295X.99.3.440

McKoon, G., & Ratcliff, R. (1994). Sentential context and on-line lexical decision. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *20*(5), 1239–1243. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7931103

McKoon, G., & Ratcliff, R. (1998). Memory-based language processing: Psycholinguistic research in the 1990s. *Annual Review of Psychology*, *49*, 25–42. doi:10.1146/annurev.psych.49.1.25
McKoon, G., & Ratcliff, R. (2013). Aging and predicting inferences: A diffusion model analysis. *Journal of Memory and Language*, *68*(3), 240–254. doi:10.1016/j.jml.2012.11.002
McKoon, G., & Ratcliff, R. (2015). Cognitive theories in discourse processing research. In E. J. O'Brien, A. E. Cook, J. Lorch, & F. Robert (Eds.), *Inferences during reading 2* (pp. 42–67). Cambridge, UK: Cambridge University Press.
McKoon, G., & Ratcliff, R. (2016). Adults with poor reading skills: How lexical knowledge interacts with scores on standardized reading comprehension tests. *Cognition*, *146*, 453–469. doi:10.1016/j.cognition.2015.10.009
McKoon, G., Ratcliff, R., & Ward, G. (1994). Testing theories of language processing: An empirical investigation of the on-line lexical decision task. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *20*(5), 1219–1228. doi:10.1037/0278-7393.20.5.1219
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. doi:10.1177/0022219409359345
Mellard, D. F., Woods, K. L., Desa, Z. D., & Vuyk, M. A. (2013, August). Underlying reading-related skills and abilities among adult learners. *Journal of Learning Disabilities*. doi:10.1177/0022219413500813
Miller, B., McCardle, P., & Hernandez, R. (2010). Advances and remaining challenges in adult literacy research. *Journal of Learning Disabilities*, *43*(2), 101–107. doi:10.1177/0022219409359341
Murray, J. D., & Burke, K. A. (2003). Activation and encoding of predictive inferences: The role of reading skill. *Discourse Processes*, *35*(2), 81–102. doi:10.1207/S15326950DP3502_1
Oakhill, J. (1982). Constructive processes in skilled and less skilled comprehenders' memory for sentences. *British Journal of Psychology (London, England: 1953)*, *73*(Pt. 1), 13–20. doi:10.1111/j.2044-8295.1982.tb01785.x
Oakhill, J. (1983). Instantiation in skilled and less skilled comprehenders. *The Quarterly Journal of Experimental Psychology*, *35*(3), 441–450. doi:10.1080/14640748308402481
Oakhill, J. V. (1984). Inferential and memory skills in children's comprehension of stories. *British Journal of Educational Psychology*, *54*(1), 31–39. doi:10.1111/j.2044-8279.1984.tb00842.x
Oakhill, J. (1993). Children's difficulties in reading comprehension. *Educational Psychology Review*, *5*(3), 223–237. doi:10.1007/BF01323045
Oakhill, J., & Yuill, N. (1986). Pronoun resolution in skilled and less-skilled comprehenders: Effects of memory load and inferential complexity. *Language and Speech*, *29*(1), 25–37. doi:10.1177/002383098602900104
Oakhill, J., Yuill, N., & Donaldson, M. L. (1990). Understanding of causal expressions in skilled and less skilled text comprehenders. *British Journal of Developmental Psychology*, *8*, 401–410. doi:10.1111/j.2044-835X.1990.tb00854.x
Oakhill, J. V., Yuill, N., & Parkin, A. (1988). Memory and inference in skilled and less-skilled comprehenders. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory* (Vol. 2, pp. 315–320). Chichester, UK: Wiley.
O'Brien, E. J. (1995). Automatic components of discourse comprehension. In R. F. Lorch & E. J. O'Brien (Eds.), *Sources of coherence in reading* (pp. 159–176). Hillsdale, NJ: Erlbaum.
O'Brien, E. J., Cook, A. E., & Lorch, R. F. J. (Eds.). (2015). *Inferences during reading*. Cambridge, UK: Cambridge University Press.
O'Brien, E. J., & Myers, J. L. (1999). Text comprehension: A view from the bottom up. In *Narrative comprehension causality and coherence essays in honor of Tom Trabasso* (pp. 35–53). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
OECD. (2013). *OECD Skills Outlook 2013: First results from the Survey of Adult Skills*. Retrieved from http://www.oecd-ilibrary.org/education/oecd-skills-outlook-2013_9789264204256-en
Peracchi, K. A., & O'Brien, E. J. (2004). Character profiles and the activation of predictive inferences. *Memory & Cognition*, *32*(7), 1044–1052. doi:10.3758/BF03196880
Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading*, *18*(1), 22–37. doi:10.1080/10888438.2013.827687
Perfetti, C. A., & Stafura, J. Z. (2015). Comprehending implicit meanings in text without making inferences. *Inferences during Reading*. doi:10.1017/9781107279186.002
Perfetti, C., Yang, C.-L., & Schmalhofer, F. (2008). Comprehension skill and word-to-text integration processes. *Applied Cognitive Psychology*, *22*, 303–318. doi:10.1002/(ISSN)1099-0720
Posner, M. I. (1978). *Chronometric explorations of mind* (p. 286). Oxford, UK: Oxford University Press. Retrieved from http://books.google.com/books?id=tQwHAAAACAAJ&printsec=frontcover\npapers2://publication/uuid/775A7210-14E1-4951-AB23-C49663FEDD23
Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. doi:10.1037/0033-295X.85.2.59
Ratcliff, R. (2002). A diffusion model account of response time and accuracy in a brightness discrimination task: Fitting real data and failing to fit fake but plausible data. *Psychonomic Bulletin & Review*, *9*(2), 278–291. doi:10.3758/BF03196283
Ratcliff, R. & Childers, R. (2015). Individual differences and fitting methods for the two-choice diffusion model. *Decision*, *2*, 237–279.

Ratcliff, R., & McKoon, G. (1988). A retrieval theory of priming in memory. *Psychological Review*, *95*(3), 385–408. doi:10.1037/0033-295X.95.3.385
Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922. doi:10.1162/neco.2008.12-06-420
Ratcliff, R., Thapar, A., & McKoon, G. (2004). A diffusion model analysis of the effects of aging on recognition memory. *Journal of Memory and Language*, *50*(4), 408–424. doi:10.1016/j.jml.2003.11.002
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*(1), 56–66. doi:10.1037/0882-7974.22.1.56
Ratcliff, R., Thapar, A., & McKoon, G. (2010). Individual differences, aging, and IQ in two-choice tasks. *Cognitive Psychology*, *60*(3), 127–157. doi:10.1016/j.cogpsych.2009.09.001
Ratcliff, R., Thapar, A., & McKoon, G. (2011). Effects of aging and IQ on item and associative memory. *Journal of Experimental Psychology. General*, *140*(3), 464–487. doi:10.1037/a0023810
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* (pp. 3–32). Oxford, UK: 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. doi:10.1016/j.cognition.2014.12.004
Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, *9*(3), 438–481. doi:10.3758/BF03196302
Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review*, *106*(2), 261–300. doi:10.1037/0033-295X.106.2.261
Sanford, A. J., & Garrod, S. C. (2005). Memory-based approaches and beyond. *Discourse Processes*, *39*(2–3), 205–224. doi:10.1080/0163853X.2005.9651680
Singer, M., Andrusiak, P., Reisdorf, P., & Black, N. L. (1992). Individual differences in bridging inference processes. *Memory & Cognition*, *20*(5), 539–548. doi:10.3758/BF03199586
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*(4), 679–688. doi:10.1037/a0037667
Todaro, S., Millis, K., & Dandotkar, S. (2010). The impact of semantic and causal relatedness and reading skill on standards of coherence. *Discourse Processes*, *47*(5), 421–446. doi:10.1080/01638530903253825
Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*(3), 550–592. doi:10.1037/0033-295X.108.3.550
van den Broek, P., Beker, K., & Oudega, M. (2015). Inference generation in text comprehension: Automatic and strategic processes in the construction of a mental representation. In E. J. O'Brien, A. E. Cook, & R. F. J. Lorch (Eds.), *Inferences during reading* (pp. 94–121). Cambridge, UK: Cambridge University Press. doi:10.1017/9781107279186.006
van den Broek, P., Rapp, D., & Kendeou, P. (2005). Integrating memory-based and constructionist processes in accounts of reading comprehension. *Discourse Processes*, *39*(2), 299–316. doi:10.1207/s15326950dp3902&3_11
van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York, NY: Academic Press.
White, C. N., Ratcliff, R., Vasey, M. W., & McKoon, G. (2010). Using diffusion models to understand clinical disorders. *Journal of Mathematical Psychology*, *54*(1), 39–52. doi:10.1016/j.jmp.2010.01.004
Whitney, P., Ritchie, B. G., & Clark, M. B. (1991). Working memory capacity and the use of elaborative inferences in text comprehension. *Discourse Processes*, *14*(2), 133–145. doi:10.1080/01638539109544779
Yuill, N., Oakhill, J., & Parkin, A. (1989). Working memory, comprehension ability and the resolution of text anomaly. *British Journal of Psychology*, *80*(3), 351–361. doi:10.1111/bjop.1989.80.issue-3

Appendix. Predictions and Data for Accuracy and Quantile Response Times

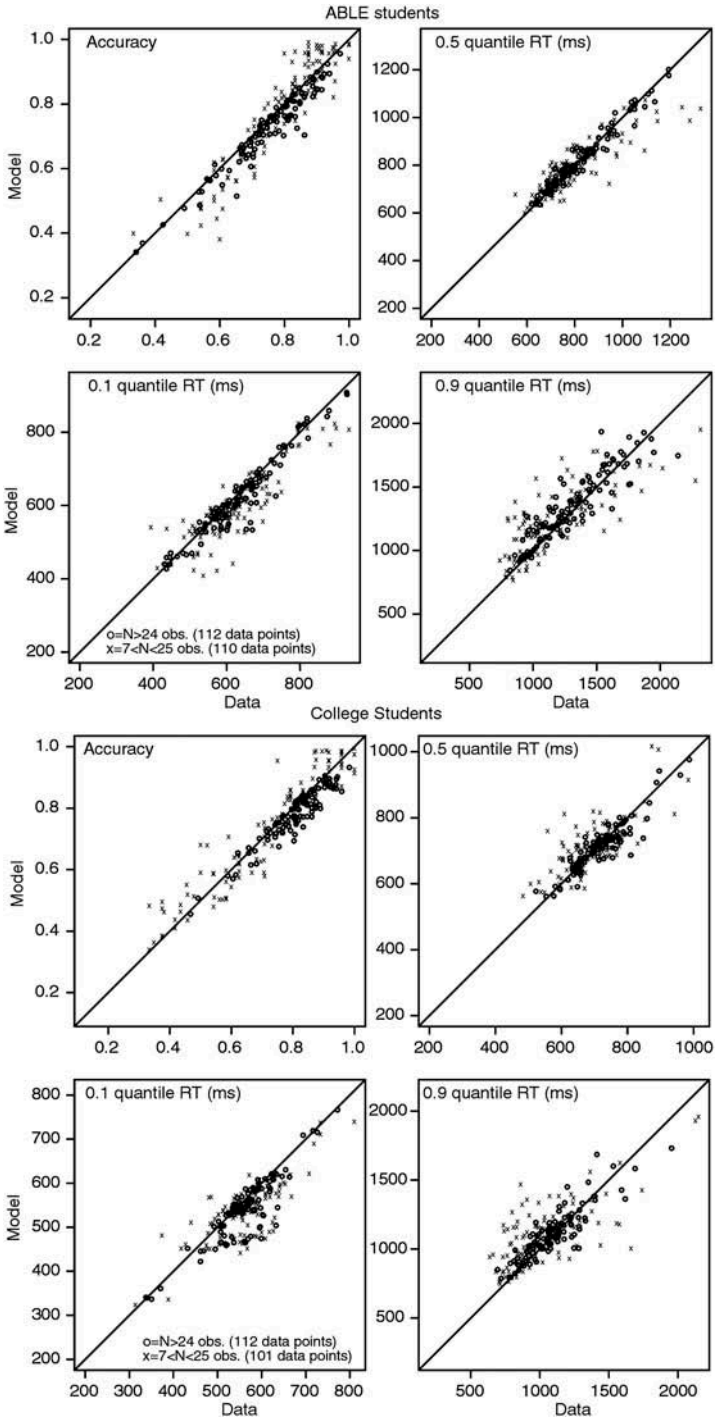

Figure A1. Plots of accuracy and the .1, .5 (median), and .9 response-time (RT) quantiles for data (x-axis) and predicted values from fits of the diffusion model (y-axis) for correct responses for college students and ABLE students.
Note. RT = response time.