Published in final edited form as:
J Speech Lang Hear Res. 2013 August ; 56(4): . doi:10.1044/1092-4388(2012/12-0145).

Test-retest reliability of eye tracking in the visual world paradigm for the study of real-time spoken word recognition

Ashley Farris-Trimble and **Bob McMurray**
Dept. of Psychology, Dept. of Communication Sciences and Disorders and Delta Center, University of Iowa

Abstract

Purpose—Researchers have begun to use eye tracking in the visual world paradigm (VWP) to study clinical differences in language processing, but the reliability of such laboratory tests has rarely been assessed. This paper assesses test-retest reliability of the VWP for spoken word recognition.

Methods—Participants performed an auditory VWP task in repeated sessions and a visual-only VWP task in a third session. We performed correlation and regression analyses on several parameters to determine which reflect reliable behavior and which are predictive of behavior in later sessions.

Results—Results showed that the fixation parameters most closely related to timing and degree of fixations were moderately-to-strongly correlated across days, while the parameters related to rate of increase or decrease of fixations to particular items were less strongly correlated. Moreover, when including factors derived from the visual-only task, the performance of the regression model was at least moderately correlated with Day 2 performance on all parameters ($R > .30$).

Conclusions—The VWP is stable enough (with some caveats) to serve as an individual measure. These findings suggest guidelines for future use of the paradigm and for areas of improvement in both methodology and analysis.

Introduction

Studies of individual differences in language often employ a variety of measures of language *performance outcomes*, to document factors that predict performance or are related to clinical diagnoses. In contrast, work in psycholinguistics focuses less on outcomes, emphasizing intermediate states of *processing* that can reveal the mechanisms underlying performance (Allopenna, Magnuson, & Tanenhaus, 1998; Altmann & Kamide, 1999; Marslen-Wilson, 1987; Spivey, 2007; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; Tanenhaus & Trueswell, 1995). Until recently, these *online* approaches have rarely been applied to individual differences or clinical populations (though see Desroches, Joanisse, & Robertson, 2006; McMurray, Samelson, Lee, & Tomblin, 2010; Nation, Marshall, & Altmann, 2003), in part because they were designed to test group-level inferences to discriminate theories of language processing, not to measure properties of individuals. In recent years, however, interest in characterizing outcome differences in terms of underlying processes has led to the use of real-time methods with clinical populations (Brock, Norbury, Einav, & Nation, 2008; Desroches et al., 2006; Farris-Trimble, McMurray, Cigrand, & Tomblin, submitted; McMurray et al., 2010; Nation et al., 2003).

Eye tracking in the visual world paradigm (VWP) has emerged as a key tool for understanding real-time language comprehension in normal listeners at levels ranging from speech perception to pragmatics (Allopenna et al., 1998; Altmann & Kamide, 1999; Hanna & Tanenhaus, 2004; McMurray, Tanenhaus, & Aslin, 2002; Sedivy, Tanenhaus, Chambers, & Carlson, 1999; Tanenhaus et al., 1995). VWP tasks are simple. An auditory stimulus instructs the listener to perform a task in a visual world, for instance to select a particular visual stimulus (from a set of computer images or real objects). The nature of the set of competing visual objects is manipulated to represent possible interpretations. Fixations to each item are recorded during the task. Because eye-movements are easy to initiate (and usually necessary for the behavioral response), listeners make fixations throughout the utterance that can reveal their preliminary commitments to various interpretations. VWP studies of spoken word recognition using phonologically related competitors (Allopenna et al., 1998; Dahan, Magnuson, & Tanenhaus, 2001; McMurray et al., 2010; McMurray, Tanenhaus, & Aslin, 2009) show a close alignment between proportion of fixations and estimates of lexical activation from continuous-time processing models like TRACE (McClelland & Elman, 1986).

The VWP has seen mounting use to examine clinical populations (Brock et al., 2008; Desroches et al., 2006; Farris-Trimble et al., submitted; McMurray et al., 2010; Nation et al., 2003). This is a natural extension. The task has low memory and executive-function demands and does not require meta-linguistic decision-making. It is also fairly ecological, tapping language skills that impaired listeners use every day, but at the same time it is sensitive to subtle differences that often cannot be observed in overt responses (McMurray, Aslin, Tanenhaus, Spivey, & Subik, 2008; McMurray, Clayards, Tanenhaus, & Aslin, 2008). In many ways, then, the VWP is an ideal tool for understanding real-time language processing in clinical populations.

However, most clinical measures describe differences between individuals or establish clinically relevant groups. For this, a test must measure some consistent property of an individual's behavior, and not be only meaningful as a group measure. Thus, the first step toward applying the VWP clinically is to determine its reliability. This is not necessarily simple. The VWP provides a rich temporal measure of multiple components of the time course of processing (e.g., the speed of fixating the right answer, the degree to which competitors are suppressed). As a result, an evaluation of reliability is inherently multi-faceted and is critical for knowing which components of a typical VWP measure are most likely to reveal subtle effects.

The first goal of this study is to examine test-retest reliability for an instantiation of the VWP that measures lexical activation and competition. As we describe, spoken word recognition was chosen for its significance to a wide variety of clinical groups. Secondarily, like all measures, the VWP is not a pure test of language; it makes use of processes like decision-making and visual search that may also be impaired in some populations. Thus, our second goal was to develop a task to estimate some of these abilities and determine how strongly they predict performance in the VWP. This task may serve as a covariate to help understand or account for these visual/cognitive differences in future studies.

Spoken Word Recognition

While the VWP has been applied to domains as diverse as low-level speech perception (McMurray, Aslin, et al., 2008; McMurray et al., 2002), word recognition (Dahan, Magnuson, & Tanenhaus, 2001; Magnuson, Dixon, Tanenhaus, & Aslin, 2007), syntactic parsing (Altmann & Kamide, 1999; Tanenhaus et al., 1995), referential interpretation (Sedivy et al., 1999) and pragmatics (Hanna & Tanenhaus, 2004), we tested the reliability of the VWP in the context of spoken word recognition. This was in part due to the number of

studies using the VWP to study word recognition in clinical populations (Brock et al., 2008; Desroches et al., 2006; Dollaghan, 1998; Farris-Trimble et al., submitted; McMurray et al., 2010; Yee & Sedivy, 2006) and in part due to the fact that speech perception (which is also of wide-spread clinical importance) is clear component of word recognition. Thus, focusing on word recognition allows us to evaluate the reliability of the paradigm for both measures. More practically, the functions describing changes in fixation behavior over time during VWP measures of spoken word recognition are well understood and can be parameterized statistically (McMurray et al., 2010; Mirman, Dixon, & Magnuson, 2008). This allows us to investigate these issues more deeply. Thus, we start with a short background on spoken word recognition before presenting the details of our own investigation.

To recognize a word, listeners must perceive and process the sounds that make up that word in real time. This is not insignificant; any given portion of the signal contains cues to multiple phonemes across the word that must be integrated. Over the past few decades, researchers have arrived at a consensus about some of the fundamental principles of this process (Gaskell & Marslen-Wilson, 1997; Marslen-Wilson, 1987; McClelland & Elman, 1986). Lexical activation is *immediate*; listeners begin to activate possible lexical candidates as soon as they receive input (Allopenna et al., 1998; Marslen-Wilson & Zwitserlood, 1989). They activate all candidates that match the received signal in *parallel* (Spivey, Grosjean, & Knoblich, 2005), and as more of the signal is received, the candidate set is updated *incrementally* so that candidates that no longer match the signal begin to be suppressed (Allopenna et al., 1998; Dahan & Gaskell, 2007; Marslen-Wilson, 1987). Candidates compete with one another: words that are more frequent or better matches to the input inhibit other less favored words (Dahan, Magnuson, Tanenhaus, & Hogan, 2001; Luce & Pisoni, 1998).

Much of our understanding of lexical processing derives from studies using online methods to tap intermediate states of processing, like cross-modal priming, gating, event-related potentials (ERP), and the VWP. In typical VWP experiments, the names of the pictures are phonologically related to the auditory stimulus. For instance, if the auditory stimulus is *horn*, a screen might contain pictures of a *horn*, a *horse*, *corn*, and a *box*. *Horn* and *horse* overlap at the onset (cohorts), while *horn* and *corn* are rhymes, and *horn* and *box* are unrelated. By measuring the proportion of looks to each item when *horn* is heard, one can estimate the relative activation of each word on a millisecond-by-millisecond time scale. In paradigms like this, listeners typically make between one and five fixations per trial. As each fixation lasts 200–300 ms, the results from any single trial are coarse and discrete. However, when these data are averaged across many trials, but within tiny time-slices, smooth functions depicting the time course of processing emerge (e.g., Figure 1). Here, fixations to the target and cohort begin to deviate from the words that do not overlap at onset roughly 200 ms after the onset of the stimulus (as it takes around 200 ms to plan and launch an eye-movement). Fixations to the cohort gradually subside after disambiguating speech material arrives, and listeners briefly consider the rhyme as its overlap with the target is heard. Such functions are characteristic of this paradigm and appear in many studies (Allopenna et al., 1998; Dahan & Gaskell, 2007; Dahan, Magnuson, & Tanenhaus, 2001; Magnuson et al., 2007; McMurray et al., 2010), and they can be characterized with non-linear functions, allowing a precise description of the behavior.

The basic principles of lexical activation have been characterized in the population of typically developing adults and children, but it is unclear whether those same principles apply to impaired populations, and researchers have been using the VWP to address. This has yielded important insights in a number of populations. For instance, Yee, Blumstein and Sedivy (2008) showed that individuals with Broca's or Wernicke's aphasia behave similarly with regards to semantic competitors, but differently in relation to phonological competitors.

The task may also reveal unexpected group differences. Brock et al. (2008) found that language impairment, not autism, differentiated children's abilities to integrate higher-level (sentential/pragmatic) information during word recognition. Desroches et al. (2006) showed that the presence of a cohort competitor slowed looks to the target equally for children with dyslexia and normal readers, but unlike normal readers, readers with dyslexia were not slowed by a rhyme. Finally, Revill and Spieler (2012) showed that older listeners fixate high-frequency phonological competitors more than younger listeners, though they fixate low-frequency competitors at the same rate, suggesting they are slightly more reliant on top-down factors. The VWP can even help pinpoint the particular portion of the time course in which impairment affects processing. For example, cochlear implant users are delayed in fixating the target and suppressing fixations to the competitor and show decreased target and increased competitor fixations even late in the time course (Farris-Trimble et al., submitted). Finally, because fixation functions can be mapped onto predictions made by TRACE, it is possible to use parameters in the model to determine the locus of the impairment. For instance, McMurray et al. (2010) showed that the activation patterns of children with language impairment suggested difficulties in lexical decay when modeled in TRACE.

The VWP has thus proven to be a useful tool in characterizing group differences in impaired populations. To look at *individual differences*, however, the validity and reliability of the task itself must be assessed. That is, to correlate an individual's performance in the VWP with individual measures like language, reading, or IQ scores, we first need to establish that the VWP measure is reliable. This is important for determining the power of a given study design and the range of such correlations one might expect. It may also help better interpret prior results—an estimate of the reliability of the measure can help determine whether weak or null effects may derive from instability in that component of the fixation record. This leads to three questions.

First is the issue of validity. Validity of an experimental measure is hard to quantify, as there are no "true" measures of theoretical constructs like lexical activation. As a group measure, the VWP mirrors many classic effects in word recognition seen with other measures (Table 1). The time course of fixations has also been shown to be highly correlated ($R^2 > .95$) with activation from computational models like TRACE (Dahan, Magnuson, & Tanenhaus, 2001; McMurray et al., 2010; McMurray et al., 2009; Spivey-Knowlton & Allopenna, 1997). Thus, at group levels the VWP achieves construct validity by capturing meaningful lexical activation/competition processes.

Second, with respect to reliability, VWP measures of spoken word recognition are multi-faceted: activation for each competitor is a continuous function of time, and we typically study multiple competitors. There are many ways to quantify meaningful properties of these functions (cf. McMurray et al., 2010; Mirman et al., 2008), but minimally we must consider reliability of this measure for multiple components of the time course (e.g., those measuring both timing and degree of fixations), and for multiple types of competitors (e.g., targets, cohorts, rhymes). Prior studies on clinical groups have all found effects on different components of the fixation curve. Desroches et al. (2006), for example, found effects that appear to be driven by the speed at which looks to the target reach asymptote[1]; McMurray et al. (2010) found them in the asymptotic fixations to targets and competitors; and Farris-Trimble et al. (submitted) found them both in the timing of peak fixations to competitors, and the asymptotic fixations to the target and competitors. Thus, we must evaluate the reliability of individual components of the fixation record.

[1]Desroches et al.'s (2006) analysis assessed the proportion of looks in a time window. However, during this window the temporal profiles they report suggest that these effects may be driven by the steep increase in looks to the target over time, which is similar to the measure of slope we describe shortly.

Finally, the VWP is not a pure measure of language. It requires memory, attention, and visuo-cognitive processes: Participants must categorize the objects, find the target, and so forth. It is possible that apparent differences in language or lexical processes derive from differences in visual categorization, visual search, or decision-making. In fact, such deficits have been documented in many groups including children with SLI (Ullman & Pierpont, 2005), hearing impairment (Smith, Quittner, Osberger, & Miyamoto, 1998), and dyslexia (Biscaldi, Gezeck, & Stuhr, 1998). Thus, it is important to develop ways to document differences in non-linguistic abilities and determine how much of the reliability of a given component of this task is due to stability in visual processes.

To address these questions, participants completed three tasks in separate sessions. Two assessed test/re-test reliability, using a basic VWP task like the *horn/horse/corn* example described above. The third was a visual version of the task developed to isolate the effects of visuo-cognitive processing. This assessed whether some variance in auditory performance is due to stable visual abilities. Moreover, accounting for visual abilities *over and above* test/retest reliability may reveal that performance on this task is more reliable than when measured in a strict test/re-test design.

Methods

The study took place over three sessions, typically one week apart. These consisted of two adjacent auditory sessions and a visual session that occurred either before or after the two auditory sessions (counterbalanced between subjects).

Participants

Twenty-six undergraduates participated for course credit. An additional eight participants were paid for their participation for a total of 34 participants. All reported normal or corrected-to-normal vision and normal hearing. Five of these participants were excluded, three because they did not participate in all three days of the experiment and two who made so few eye movements that their data were unanalyzable. The data from the remaining twenty-nine subjects were analyzed.

Auditory/Lexical Task

Design—Forty sets of four words were selected (Supplemental Material Appendix 1). Each contained a target (e.g., *horn*), a cohort (e.g., *horse*), a rhyme (e.g., *corn*), and an unrelated word (e.g., *box*). All words were picturable nouns and within a set were semantically unrelated.

On each trial, pictures corresponding to the four words of a set appeared, and one word was played. Each word served as the auditory stimulus four times yielding 640 trials (40 sets × 4 words/set × 4 repetitions). There were four trial-types, defined by which word was the auditory stimulus. For example, in the *horn/horse/corn/box* set, when *horn* was the stimulus, *horse* and *corn* were competitors (a cohort and rhyme). When *corn* was played, though, only *horn* (a rhyme) was a competitor—the other two words were unrelated. Table 2a illustrates the roles played by each word in a set as a function of trial-type. (The letters naming each trial type refer to the roles the objects played in that particular trial). We estimated functions for competitor fixations across multiple trial-types: target fixations, for example, can be examined on all four trial-types, cohort fixations on TCRU and TCUU trials, and so on. Each trial-type was one fourth of the 640 trials.

Auditory stimuli—The 160 words were recorded by a male speaker of Midwestern American English who had been trained in articulation. Recordings were made in an

anechoic chamber at a sampling rate of 44, 100Hz on a Marantz PMD670. Each word was recorded 4–5 times in the carrier phrase "He said ___." From these, the clearest token was selected and excised from the carrier sentence. One-hundred milliseconds of silence were added to the onset of each sound-file.

Visual stimuli—For each of the 160 stimuli, an easily recognizable clipart-style picture was developed in a selection process used in prior studies (Apfelbaum, Blumstein, & McMurray, 2011; McMurray et al., 2010; Toscano & McMurray, 2012). For each word, multiple candidate pictures were obtained from a commercial clipart database. A committee of 3–4 people then selected the picture that best corresponded to that word. Pictures were edited as needed to remove extraneous elements, change colors, and to ensure that the pictures had relatively equivalent visual salience. Finally, the pictures were approved by a lab member with extensive experience in the VWP.

Procedure—The experiment was programmed with Experiment Builder software (SR Research Ltd., Ontario, Canada). Participants were seated in front of a 1280×1024 computer monitor wearing Sennheiser HD 280pro headphones. Eye-movements were tracked with a table-mounted camera (Eyelink I000, SR Research Ltd.), and a chin rest was used to stabilize the head. The eye-tracker was calibrated at the beginning of the experiment using a standard 9-point procedure. Optional breaks and a mandatory drift correction procedure occurred every 32 trials. The experiment lasted 45–50 minutes (including eye tracker setup).

Supplemental Material Figure S1a shows a sample display from the task. On each trial, the four pictures from a set were displayed at 300×300 pixels in the corners of the screen with a small red circle in the center. Each picture was 50 pixels from the nearest edge. After 500 ms (so the participants could view the pictures), the circle turned blue, the participant clicked on it and heard the auditory stimulus. They then clicked on the corresponding picture. Pictures were randomly assigned to the four corners on each trial. While the same stimuli were used in both sessions, presentation order and picture locations were randomized independently for each session. Though using the same stimuli across days can create practice effects, this was done to provide the strongest measure of test-retest reliability.

Visual Task

Design—A visual analogue of the auditory task was developed to estimate the contribution of visual processes. Rather than matching an auditory word to phonologically related items, participants matched a briefly presented colored shape to visually related items. Each set of four pictures consisted of four types of colored shapes: a target (e.g., a red triangle), a color-match (e.g., a red circle), a shape-match (e.g., a blue triangle), and an unrelated picture (e.g., a yellow square) (Supplemental Material Appendix 2). The target stimulus appeared in the middle of the screen, and participants clicked the matching shape. As in the auditory task, each shape in a set occurred as the target four times creating four trial-types (Table 2b) and a total of 640 trials.

Stimuli—160 colored shapes were created from a possible set of 12 colors and 16 basic shapes (square, circle, triangle, etc.). The colored shapes were arranged in sets of four, as described above (Supplemental Material Figure S1b). Shapes were 300×300 pixels.

Procedure—The procedure was similar to the auditory task. The four shapes in a set were presented at the same size and locations as the pictures in the auditory task. After the participant clicked on the central fixation circle, the target shape (also 300×300 pixels) appeared in the center for 100 ms. Because the 100 ms target presentation was shorter than

the length of an average word, this session tended to be a little shorter (35–40 minutes) than the auditory sessions.

Eye tracking

A table-mounted Eyelink I000 (SR Research Ltd.) tracked participants' eye movements at 250 Hz. The fixation record was automatically parsed into fixations, saccades and blinks using the default "psychophysical" parameter set. As in prior studies, we combined saccades and the subsequent fixation into a single "look." When assigning looks to pictures, the borders were extended by 100 pixels in each direction (to account for minor drift in the tracker). This did not result in any overlap. On average, it takes around 200ms to plan and launch an eye movement (Matin, Shao, & Boff, 1993). Our auditory stimuli included 100ms of silence before the onset of the word. We thus ignored eye-movements launched during the first 300ms of each trial as not reflective of a response to the stimulus.

Results

Our analyses addressed five questions. First, we examined overall performance (both mouse-click accuracy and fixations). Next, we performed a gross correlational analysis on the Day 1 and Day 2 auditory tasks to determine which portions of the time course were most reliable. Third, we examined specific parametric components of the fixation record to calculate the reliability of meaningful parameters that describe the time course of processing. Fourth, we conducted hierarchical regression analyses to determine the separate contributions of the visual and auditory tasks and to examine extraneous factors (e.g., *when* the visual task was performed). Finally, we compared the two auditory tasks to see if performance changed across days.

Overall Performance

Participants were highly accurate on the auditory tasks (both days M> 99.5 percent). Response time was faster on Day 2 (Day 1: M=1422 ms; Day 2 M=1336 ms, $t(28)=6.9$, $p<0.001$).

Figure 1 shows the proportion of fixations over time to each competitor in the auditory tasks (averaged across both days) for the TCRU trials. Allowing for the 100 ms of silence before the onset of the auditory file plus the 200 ms oculomotor planning time, fixation measures start at 300 ms. Curves like these can show gross patterns in the rate, timing, and degree of fixations to assess performance on the two auditory tasks. These curves show the predicted pattern of performance, with target looks increasing at a relatively fast rate to an upper asymptote of around 87 percent. Cohort looks follow the same initial pattern but diverge around 500 ms, which is about 200 ms after the auditory material in the target word becomes inconsistent with the cohort. Rhyme fixations are fewer than cohort fixations in general, but late in the trial (after participants have heard the latter portions of the word) surpass cohort fixations.

Reliability 1: Raw correlations

To examine the gross reliability of the fixation functions, we conducted a simple correlation analysis. This focused on the test-retest reliability of looks to each item type at each time slice. Given the complexity of the time course and the fact that various studies have found different effects on several distinct portions of it, we wanted to determine (roughly) the relative reliability of different portions of the time course for each class of lexical competitors. Thus, for each of the four types of competitors, we computed the pairwise correlation between the proportion of fixations on each time slice on Day 1 with the corresponding time slice on Day 2.

We do not make any claims about the significance of the correlations at each point, as it is unclear what family-wise alpha should be. There are a large number of comparisons, and these correlation estimates are not independent, as any given fixation spans several hundred milliseconds (i.e. 20–60 data points). Thus, as we wanted to start with a coarser level of analysis that does not assume any particular analytic approach, we use R here as an index of reliability, because this approach does not make any strong assumptions about the shape of the function (as will the more precise analyses presented shortly).

Figure 2 shows the results, overlaying the correlations at each time on the average fixation curves. For the target we found moderate or better correlations[2] at each time slice of target fixations (Figure 2a). The strongest correlations (up to R≈.79) occurred in the early parts of the trial, as fixations were building. Correlations dropped somewhat at the asymptote (to R≈.39), suggesting that fixations late in the trial are more variable from day-to-day within a participant. The correlations for competitors show an interesting discrepancy. For both the cohort (Figure 2b) and the rhyme (Figure 2c), Day 1 and Day 2 fixations were highly correlated at the time at which competitor looks reach their peak and weaken as participants' fixations to those items diminish. Cohort looks, however, remained moderately correlated for the remainder of the time course, averaging R = .47 between 1000 and 2000 ms, while the rhyme correlations dropped dramatically (R < .2) and then rise again to an average R = .66 between 1300 and 2000 ms. This is particularly interesting, as there are many fewer looks to the rhyme overall. This result implies that while late looks to the rhyme may be minimal, they are highly consistent. Finally, fixations to the unrelated item (Figure 2d) are moderately correlated throughout the time course, without much fluctuation.

While this analysis offers a broad overview of how reliable participants' behavior is over time, it obscures between-participant variability. Some participants reach their peak looks to the competitor earlier than others; some participants reach a greater asymptote in target fixations than others, etc. To better characterize such differences, one approach is to estimate specific properties of the time course, and then compare matching components of the function (e.g., the time of peak fixations), rather than comparing the functions at matching times. Thus, the next section presents an analysis based on McMurray et al. (2010) that identifies some meaningful landmarks, extracts numeric estimates, and correlates these measures across days.

Reliability 2: A parametric approach

Statistical Methods—We characterized each participant's fixation record as a mathematical function of time and used the parameters of this function to quantify the time course. One approach to doing this is to use polynomial functions to describe the fixation curves (Mirman et al., 2008). These are easy to estimate and can be integrated with linear growth models. However, the parameters of polynomial functions do not map in a one-to-one manner onto what are likely to be meaningful properties of the time course (e.g., there is no parameter for the asymptote or timing of the peak). Moreover, polynomial functions can only approximate these functions; for example, they cannot asymptote the way the target curve does, without including fairly high (6^{th} or 7^{th}) order terms. This requires more parameters and makes interpretation more difficult. Thus, following McMurray et al. (2010), we used nonlinear functions. Because target fixations build gradually, then increase exponentially and asymptote, we fit a logistic function to this data. In contrast, because competitor fixations build, peak, and then decrease, these better corresponded to a Gaussian-like function.

[2]We take correlations between .10 and .29 as small effects, between .30 and .49 as moderate effects, and .50 or greater as large effects (Cohen & Cohen, 1983).

For each participant on each day, looks to the target (the auditory stimulus) as a function of time were averaged across all correct trials. Trials ended when the participant selected a picture, and as a result there was variation in the duration of the trial both within and between subjects. To cope with this, each participant's data were truncated at their mean reaction time, and for trials that were shorter than this RT, the duration of the final eye-movement was extended, following McMurray et al. (2010).[3] We chose to constrain the data in this way for several reasons. First, we make the basic assumption that each participant's RT represents the time by which looks to the target have reached asymptote; looks after that may not be driven by the auditory stimulus. Second, had we chosen a single arbitrary cutoff for all subjects, the bulk of some subjects' trials would have been lengthened, while others' would have been shortened. Our method results in an approximately equal amount of shortened and lengthened trials for each subject (see also McMurray et al., 2010). Finally, because we are seeking measures of individual performance, rather than the coarser overall measures in the raw correlation analysis above, this method avoids a situation in which a participant's data is influenced by other participants (e.g., as would be the case if we chose a single average cutoff for all subjects).

A logistic function (Figure 3A, Equation 1) was fit to the time course of target fixations.

$$P(\text{target})\frac{p-b}{1+\exp\left(4\cdot\frac{s}{p-b}\cdot(c-t)\right)}+b \quad (1)$$

This function has four parameters: the baseline (*b*), corresponding to the starting asymptote; the peak (*p*), corresponding to the final asymptote; the crossover (*c*), or the point in time at which the function transitioned; and the slope (*s*), the derivative of the function at the midpoint. These four parameters were estimated for each participant and each day.

The shape of the curves for competitor fixations corresponds to something resembling a Gaussian. We used a function created for this type of analysis (see McMurray et al., 2010) that allows the left and right sides of the function to vary independently (Figure 3B, Equation 2).

$$P(competitor)=\begin{cases}\exp\left(\frac{(t-\mu)^2}{-2{\sigma_1}^2}\right)(p-b_1)+b_1 & \text{if } t\leq\mu\\ \exp\left(\frac{(t-\mu)^2}{-2{\sigma_2}^2}\right)(p-b_2)+b_2 & \text{if } t>\mu\end{cases} \quad (3)$$

This function has six parameters: onset baseline (b_1), onset slope (σ_1), midpoint (μ), peak height (*p*), offset slope (σ_2), and offset baseline (b_2). Onset and offset baselines are measures of the initial and final proportion of fixations to these items. The onset and offset slopes correspond with the rate of increase and decrease of fixations to the items. The midpoint is the location (in time) when fixations to that object peak, and the peak height is the proportion of fixations at this time.

Both functions were fit using a constrained gradient descent method that minimized the root-mean-squared error between the function and the data (Matlab code available from the second author). Multiple starting parameters were used, and the final fits were inspected visually to ensure that the fits were not a local minimum. The results of this were a set of parameters that describe meaningful properties of each participant's time course of fixations

[3]The average standard deviation for RTs across all experimental days was 263 ms, illustrating that the amount of variance in RT was not large, and the lengthening or truncation was thus typically small.

to each competitor, both in terms of timing (slopes, midpoint and crossover point) and degree (baseline and maximum/peak fixations). These were used to determine the reliability of each estimate. Because the first 300ms of each trial are set to zero, the lower asymptotes of both curves are not informative, and the onset slope is less free to vary. We thus excluded the onset baseline measurements from the following analyses.

Reliability—Using this more parametric description of the time course of fixations, we assessed reliability by correlating the Day 1 estimate of each parameter with its Day 2 estimate. It is important to note that a simple correlation analysis likely leaves a lot of variance unexplained (e.g., whether the visual task was performed before or after the two auditory tasks). These sources of variance will be assessed in the next section. However, we report raw correlations here because this is the standard by which many clinical assessments are judged.

The top of Table 3 shows correlations across days for each of the target logistic parameters. All were significantly correlated. The *timing* parameters (slope and crossover) were highly correlated, while the asymptote (*degree*) showed a moderate correlation. The lower sections of Table 3 show analogous results for competitors. Here, the midpoint and peak height were highly reliable. The offset baseline was moderately-to-strongly reliable for the cohort and rhyme, but not for the unrelated item. Finally, the cohort onset slope and rhyme offset slope were moderately reliable, while the inverse parameters (cohort offset slope and rhyme onset slope) were not. This finding is noteworthy because the more highly correlated slope parameters are the ones that line up with the phonetic overlap between target and competitor: cohorts overlap with the target early, and the cohort *onset* slope was moderately reliable; while the rhymes overlap later, and rhyme *offset* slope was also moderately reliable. This dissociation may indicate that the degree to which a particular component of the fixation function is driven by the auditory signal is a strong contributor to the reliability of behavior at that time.

In sum, most measures of target and competitor activation were moderately or strongly correlated across days. Onset and offset slope competitor measures were least reliable, suggesting that the rate of activation/suppression may be less stable than the timing and amount of activation.

Multiple Regression Analysis

The correlational analysis suggested that many of the parameters describing the time course of fixations were reliable. However, these measures may have been influenced by other variables, which could lead us to underestimate reliability. Most importantly, some participants received visual testing before the first day of auditory testing and others received it after the second day of auditory testing. Participants who performed visual testing first may have gained some experience with tasks of this nature, and this could have affected performance. Accounting for this main effect, as well as its interaction with Day 1 performance in predicting Day 2 performance, is thus important for accurately gauging reliability.

We performed a hierarchical regression to determine the influence of these factors and whether the Day 1 auditory task is informative in predicting outcomes on the Day 2 task above and beyond these factors. To assess visual attention and search strategies, we performed the same curvefitting analysis on fixations from the visual-only task. We used target fixations in the visual task as predictors of target fixations in the auditory task, and unrelated fixations in the visual task as predictors of unrelated fixations in the target task. Because it is not clear how the color-match and shape-match object on the visual task

correspond to the cohort and rhyme in the auditory task, both were used as predictors for cohort and rhymes.

We performed two series of regressions for each of the 18 curvefit parameters. In each, the parameter from the second auditory task was the dependent variable. The first series of regressions was used to characterize the variation, asking whether the joint combination of auditory and visual predictors increased the reliability estimates obtained above and if there were interactions between auditory Day 1 performance and visual performance in predicting Day 2. The second series used a similar framework to determine the *independent* contributions of auditory and visual predictors.

Joint contributions of auditory and visual predictors—We first sought to determine the joint contributions of auditory and visual predictors to performance on the Day 2 auditory task and to investigate interactions. The first predictor entered into this model was a nominal variable (dummy coded) indicating whether the visual task was performed before or after the first auditory task (*Vday*). A main effect of *Vday* would reflect differences in performance due to having an additional day of experience. Second, the auditory (*A1*) and visual (*V*) correlates of the dependent variable were entered together. This allowed us to determine global predictability of Day 2 performance from similar tasks, and including the visual factors may increase the reliability estimates computed above. Third, an interaction term examining the interaction of *Vday* and auditory performance was entered (*Vday x A1*) to test differences in the reliability of the auditory task due to prior experience.

Table 4 summarizes the results. Whether the visual task was administered before or after both auditory tasks (*Vday*) was only a significant predictor of rhyme onset slope and a marginally significant predictor of unrelated onset slope. Other than these factors, this variable never accounted for more than 8 percent of the variance. In most cases, the largest proportion of the variance was accounted for in step 2, where auditory and visual predictors were added. These parameters accounted for a significant proportion of the variance in all of the target parameters, in the height parameter for all competitors, the cohort offset baseline (marginally), the rhyme midpoint, and the unrelated onset slope and midpoint.

The interaction between *Vday* and the auditory parameters accounted for a significant proportion of the variance for only a handful of parameters: cohort offset slope and offset baseline (both marginally); rhyme offset slope; and unrelated onset baseline, midpoint, and offset slope. This suggests that the raw test-retest reliabilities may have under-estimated the reliability of these components, as our manipulation may have had unexpected effects on their predictability.

The last column in Table 4 shows the updated reliability after we account for visual predictors and the effect of task order, giving us a measure of the reliability that includes known variation due to the timing of the visual task and visual factors in general. Comparing this to the raw reliabilities in the correlational analysis (repeated in the penultimate column of Table 4) we see that a number of parameters that had only small or moderate effects are now quite robust. The R for the target asymptote, for example, increases from .39 to .678. In sum, behavior on other similar tasks was a strong predictor of behavior on the Day 2 auditory task, though order of task completion was not.

Independent contributions of auditory and visual predictors—In the previous regressions, the auditory and visual predictors were entered together to determine their joint influence. Here we investigate their individual impacts. We performed two more sets of regression analyses in which we split step 2 to perform a communality analysis. In one analysis, the auditory parameter was entered before the visual parameter(s), and in the other

analysis, the order was reversed. Interaction terms were dropped, as they were entered after the main effects (and would therefore have no bearing on this analysis). From these two new analyses we computed the unique variance associated with each factor. The unique auditory contribution is the proportion of the variance accounted for by the auditory parameter above and beyond the visual parameter (i.e., when the auditory parameter was entered after the visual parameter); the unique visual contribution is the reverse. The shared variance is the variance remaining after the unique auditory and unique visual contributions are removed from the total variance.

We should note that in a handful of cases (indicated in Table 5 with ^{o}), the visual predictor was *negatively* correlated with either the auditory Day 2 dependent variable or the auditory Day 1 predictor. This was significant in some cases and we have no explanation. Because of these inconsistencies, the shared variance estimates were anomalous (less than 0). These out-of-range correlations were also observed in structural equation models that were sensitive to the full pattern of covariance among the three variables. We thus do not report shared variance for these analyses.

Table 5 and Figure 4 show the variance accounted for by the auditory and visual parameters as well as their shared variance. The Day 1 auditory parameters account for a significant proportion of the variance above and beyond the visual parameters in the target slope and crossover and the midpoint and height measures for all non-target items. Moreover, Day 1 auditory cohort and unrelated onset slope, cohort and rhyme offset baseline, and rhyme offset slope are each significantly or marginally predictive of Day 2 performance beyond the visual predictors. The visual parameters account for a significant proportion of the variance above and beyond the auditory predictors in the target slope and asymptote, cohort height, rhyme height, onset slope and (marginally) cohort onset baseline and unrelated height.

Looking across parameters, we see a number of commonalities. First, within each of the sub-panels of Figure 4, the parameters are ordered by when they occur in time. Later parameters (e.g., target asymptote, cohort B2, rhyme B2) show less predictability than earlier ones. This suggests that the VWP may show the most reliability during times close to the auditory stimulus. Second, the visual and shared variance made the largest contributions to parameters describing *how much* people fixate (target asymptote and cohort, rhyme and unrelated height). Work examining these factors (e.g., McMurray et al., 2010) should attempt to estimate (and co-vary out) visual/cognitive performance. Third, the three competitors (Figure 4B–D), behave similarly—their midpoints are mostly auditory variance, with some shared component; height and B2 are more evenly split. These commonalities suggest that our findings are likely to generalize to other sorts of competitors and reflect a standard process of mapping briefly activated competitors (of many sorts) to visual referents. The exception to this is the slope parameters for the cohort and rhyme (Figure 4E) where we saw an interesting pattern; for cohorts, onset slope but not offset slope were heavily driven by the auditory task, while for rhymes this was reversed. Again, this may be due to differences in where cohorts and rhymes overlap with the auditory input.

Intriguingly, much of the variance in unrelated fixations was predicted by auditory and not visual factors. This was unexpected, as unrelated items are often included in VWP designs as a baseline measure of fixations unconnected to the auditory signal. The fact that visual factors are not strong predictors of unrelated looks could be because unrelated colored shapes simply do not tap the same types of processes as pictures. Alternatively, however, work with the TRACE model of spoken word recognition (McClelland & Elman, 1986), suggests that a number of language-relevant parameters affect unrelated activation (McMurray et al., 2010); thus, unrelated looks may reflect something general about language processing.

In sum, our second regression analysis indicates that much of the variance in the parametric components of the fixation functions can be predicted from previous performance on the same task, and that adding factors based on visual ability improves predictability, but typically not substantially. Most of the important components of the time course function were uniquely predicted by the auditory task (over and above the visual task). Thus, even though the VWP is based on visual behavior, it is clearly tapping *language* (or more precisely, word recognition), and partialling out variance related to these visual/cognitive factors may help reveal this. Taken as a whole, these results suggest that not only are our measures in large part reliable, but that a more comprehensive view of an individual's abilities (auditory and visual) may improve our assessment—an important advantage if these measures may eventually be used diagnostically.

Comparison of Day 1 and Day 2

Our final question is whether repeated testing changes performance. As a number of studies have adopted two-day designs to increase power (McMurray, Clayards, et al., 2008; McMurray et al., 2002, 2009), we wanted to determine whether the second day of testing differs systematically from the first. Figure 5 shows looks to each auditory item type on each day averaged across participants. It suggests that participants looked to the target somewhat faster on Day 2, and that all competitor items received more looks on Day 1 than on Day 2.

To examine this statistically, we used our parameterization of the time course to examine differences in performance on the Day 1 and Day 2 auditory tasks. We thus conducted pairwise t-tests on each parameter from the curvefits. Results are shown in Table 6. The rate of increase in fixations to the target (slope) was greater on Day 2 than Day 1, suggesting that participants may have been able to orient to the correct picture and/or recognize the word faster after a day's practice. However, the maximum degree of looks and the timing of those looks did not differ significantly. For all of the non-target items, participants made fewer looks at the peak and at the end of the trial (B2) on Day 2. As this also includes the unrelated item, this indicates people may have simply been engaging in less visual search as they better recognized the pictures with a day of practice. Finally, the growth in looks to the rhyme (onset slope) was slower on Day 2, and the decrease in looks to the cohort and unrelated items (offset slope) was marginally slower on Day 2. In sum, these results suggest learning: participants learned enough about the pictures that they were quicker to look to the target and less likely to look at competitor items on Day 2.

General discussion

Our analyses revealed several critical findings. First, fixations to the target item were highly reliable. All three target parameters (slope, crossover, and asymptote) were significantly correlated across days; when we include the other parameters, we can account for 46–74 percent of the variance in performance on the Day 2 task. The asymptote is the least reliable of the three target parameters, which may be because looks at the end of the time course are more likely to vary as participants select the correct object and then look away.

For the competitors, some parameters were consistent across days and others less so. The midpoint and peak height were reliable for all three competitors, indicating that the timing and degree of maximum fixations to those items is stable. The offset baseline was reliable in cohorts and rhymes, though not unrelated items—this may be because looks to the phonological competitors at the end of the trial are more likely to be governed by sustained competitor activation. When we add visual factors and task order, however, we find strong correlations (greater than R = .5) in all competitor parameters except the cohort onset and offset slopes (R = .43 and .49, respectively), and the unrelated offset slope and baseline (R = .46 and .33, respectively).

The least reliable components were the competitor onset and offset slopes, denoting the increase and decrease of fixations to competitors. It is possible that our curve-fitting methodology does a poorer job of fitting the slope portion of the function, and so our fits may be less accurate. For example, an increased offset slope could be also approximated by changing the offset baseline. However, the raw correlation analyses (which do not use the curvefitting), show a similar pattern with marked drops in reliability for rhymes and cohorts just as fixations began declining from the peak. We suspect that this may derive from the fact that rapid changes in underlying activation (increases or decreases) may not always be well reflected in the slower oculomotor system. However, more broadly, future work should use this as guidance—findings based on the absolute timing and degree of activation for competitors may be more robust, and if the interest is in how competitors are activated or deactivated, more power may be needed.

A limitation of the current study is its specificity to both task (VWP) and area of interest (word recognition). The VWP has been used to study language processing at a variety of levels, from speech perception to semantics and pragmatics, and our findings do not necessarily extend to all levels. We found reliable behavior in specific components of the time course, components that correspond to meaningful aspects of word recognition. Which components are meaningful at other levels of language may differ. Similarly, though our results indicate that word recognition can be reliably measured with the VWP, it is not clear whether other behavioral measures of word recognition (e.g., gating, priming) show similar reliability. In sum, our findings suggest that the tested paradigm is stable, but this conclusion may not be generalizable to other word recognition methods or to VWP tasks at other levels of language. However, we see no reason why other experimental designs using the VWP to study aspects of word recognition (e.g., Dahan, Magnuson, & Tanenhaus, 2001; McMurray et al., 2010) would not show similar levels of reliability.

How do our results impact other ways to analyze data from VWP studies? Many studies focus on the proportion of fixations made by a group or individual over a time-window, the “area under the curve” (Brock et al., 2008; Nation et al., 2003; Trueswell, Sekerina, Hill, & Logrip, 1999; Yee et al., 2008). We did not test the reliability of such measurements as the time-windows vary substantially between studies, and those approaches have been criticized for ignoring the detailed time course. However, our coarser correlation analysis (Figure 2) may help interpret existing studies or plan such designs in the future by aligning the time-window of interest with the data in Figure 2.

As described above, however, several studies have taken approaches similar to ours, analyzing specific components of the fixation function (McMurray et al., 2010; Farris-Trimble et al., submitted). Some of the differences between impaired and normal populations noted in these studies were in areas that we found to be reliable, like target fixations or competitor midpoint and offset baseline. However, these studies also found significant group differences in some measures were less reliable, like the competitor onset and offset slopes. This does not mean that such findings are less trustworthy. In some ways, the fact that these studies found group-wise differences in these parameters *despite* fairly low reliability suggests that they may be rather large underlying effects. Thus, while we may not wish to make strong claims about the individual participants in these studies on the basis of their scores, the group-wise effects are not invalidated.

It is also likely that we are underestimating the reliability of these specific parameters. Our participants were a fairly homogeneous group, psychology students at the University of Iowa. If parameters like the target asymptote are primarily affected by language ability, our dataset may not have much variability in this (since participants likely have similar ability levels). The greater variability in special populations could *increase* the reliability of our

measurements, if participants' language processing behaviors are atypical, but consistently and predictably so. Thus, if these particular parameters are to be used to make claims about individuals, the clear next step is to assess the reliability of the paradigm in a population with more variability.

Our results suggest that an individual's behavior in the VWP is generally stable across time, and the studies mentioned above establish that the paradigm can reveal differences between clinical populations. However, the question remains as to whether the paradigm is reliable enough to justify its use in the characterization or diagnosis of individuals within clinical populations. The reliability of laboratory tasks is rarely assessed, and so determining an "acceptable" level of reliability for use as a clinical measure is difficult. Standards of acceptable reliability depend on the function of the assessment, i.e. whether it will be used for individual diagnosis, to characterize individual differences, or to characterize group differences. On the one hand, psychometric instruments typically have reliabilities greater than .80, but it is not uncommon to find clinical measures below this. The reliability of the popular *Clinical Evaluation of Language Fundamentals-IV* (CELF-IV) ranges from .71 to .86 for subtests (Semel, Wiig, & Secord, 2003), and the *Comprehensive Assessment of Spoken Language* (CASL) has reliabilities as low as .65 (Carrow-Woolfolk, 1999). Some of the components we have measured here show reliabilities well into this range, like target slope and crossover and the midpoint and peak of the cohort and rhyme. Using a more complete assessment that includes visual processes can improve these even more, raising the possibility that these components could contribute to diagnosis down the road. In sum, then, the VWP as a measure of word recognition is both sensitive enough to measure group differences and stable enough (with some caveats) to serve as an individual measure. Future work can and should focus on determining which aspects of the fixation time course are related to differences in language ability, and which may be relevant for clinical diagnostics.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by grants from the National Institutes of Health: DC000242, DC008089, and DC011669. Many thanks to Bruce Tomblin for help discussions about the issues raised by reliability, Kristian Markon for statistical assistance with the communality analyses, Dan McEchron for administrative support and assistance with recruiting, and the members of the MACLab at the University of Iowa for helpful discussions. Thanks also to our participants.

References

Allopenna P, Magnuson JS, Tanenhaus MK. Tracking the time course of spoken word recognition using eye-movements: evidence for continuous mapping models. Journal of Memory and Language. 1998; 38(4):419–439.

Altmann GTM, Kamide Y. Incremental interpretation at verbs: Restricting the domain of subsequent reference. Cognition. 1999; 73(3):247–264. [PubMed: 10585516]

Andruski JE, Blumstein SE, Burton MW. The effect of subphonetic differences on lexical access. Cognition. 1994; 52:163–187. [PubMed: 7956004]

Apfelbaum K, Blumstein SE, McMurray B. Semantic priming is affected by real-time phonological competition: Evidence for continuous cascading systems. Psychonomic Bulletin and Review. 2011; 18(1):141–149. [PubMed: 21327343]

Biscaldi M, Gezeck S, Stuhr V. Poor saccadic control correlates with dyslexia. Neuropsychologia. 1998; 36(11):1189–1202. [PubMed: 9842764]

Brock J, Norbury C, Einav S, Nation K. Do individuals with autism process words in context? Evidence from language-mediated eye-movements. Cognition. 2008; 108(3):896–904. [PubMed: 18692181]

Carrow-Woolfolk, E. Comprehensive Assessment of Spoken Language. Circle Pines, MN: American Guidance Service; 1999.

Cohen, J.; Cohen, P. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. 2. Hillsdale, NJ: Erlbaum; 1983.

Connine CM, Blasko DG, Titone D. Do the beginnings of spoken words have a special status in auditory word recognition? Journal of Memory and Language. 1993; 32:193–210.

Connolly JF, Phillips NA. Event-related potential components reflect phonological and semantic processing of the terminal word of spoken sentences. Journal of Cognitive Neuroscience. 1994; 6:256–266. [PubMed: 23964975]

Dahan D, Gaskell MG. The temporal dynamics of ambiguity resolution: Evidence from spoken-word recognition. Journal of Memory and Language. 2007; 57:483–501. [PubMed: 18071581]

Dahan D, Magnuson JS, Tanenhaus MK. Time course of frequency effects in spoken-word recognition: Evidence from eye movements. Cognitive Psychology. 2001; 42:317–367. [PubMed: 11368527]

Dahan D, Magnuson JS, Tanenhaus MK, Hogan E. Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. Language and Cognitive Processes. 2001; 16(5/6):507–534.

Desroches AS, Joanisse MF, Robertson EK. Phonological deficits in dyslexic children revealed by eyetacking. Cognition. 2006; 100:B32–B42. [PubMed: 16288732]

Desroches AS, Newman RL, Joanisse MF. Investigating the time course of spoken word recognition: Electrophysiological evidence for the influences of phonological similarity. Journal of Cognitive Neuroscience. 2008; 21:1893–1906. [PubMed: 18855555]

Dollaghan C. Spoken word recognition in children with and without specific language impairment. Applied Psycholinguistics. 1998; 19:193–207.

Farris-Trimble, A.; McMurray, B.; Cigrand, N.; Tomblin, JB. The process of spoken word recognition in the face of signal degradation: Cochlear implant users and normal-hearing listeners. (submitted)

Gaskell MG, Marslen-Wilson W. Integrating form and meaning: a distributed model of speech perception. Language and Cognitive Processes. 1997; 12(5/6):613–656.

Gow DW, Gordon PC. Lexical and prelexical influences on word segmentation: Evidence from priming. Journal of Experimental Psychology: Human Perception and Performance. 1995; 21(2): 344–359. [PubMed: 7714476]

Hanna JE, Tanenhaus MK. Pragmatic effects on reference resolution in a collaborative task: evidence from eye movements. Cognitive Science. 2004; 28:105–115.

Kutas M, Hillyard SA. Brain potentials during reading reflect word expectancy and semantic association. Nature. 1984; 307:161–163. [PubMed: 6690995]

Luce PA, Pisoni DB. Recognizing spoken words: The neighborhood activation model. Ear and Hearing. 1998; 19(1):1–36. [PubMed: 9504270]

Magnuson JS, Dixon J, Tanenhaus MK, Aslin RN. The dynamics of lexical competition during spoken word recognition. Cognitive Science. 2007; 31:1–24. [PubMed: 21635286]

Marslen-Wilson W. Functional parallelism in spoken word recognition. Cognition. 1987; 25:71–102. [PubMed: 3581730]

Marslen-Wilson WD, Moss HE, van Halen S. Perceptual distance and competition in lexical access. Journal of Experimental Psychology: Human Perception and Performance. 1996; 22:1376–1392. [PubMed: 8953227]

Marslen-Wilson WD, Zwitserlood P. Accessing spoken words: The importance of word onsets. Journal of Experimental Psychology: Human Perception and Performance. 1989; 15:576–585.

Matin E, Shao KC, Boff KR. Saccadic overhead: Information-processing time with and without saccades. Perception and Psychophysics. 1993; 53(4):372–380. [PubMed: 8483701]

McClelland JL, Elman JL. The TRACE model of speech perception. Cognitive Psychology. 1986; 18(1):1–86. [PubMed: 3753912]

McMurray B, Aslin RN, Tanenhaus MK, Spivey MJ, Subik D. Gradient sensitivity to within-category variation in speech: Implications for categorical perception. Journal of Experimental Psychology, Human Perception and Performance. 2008; 34(6):1609–1631. [PubMed: 19045996]

McMurray B, Clayards M, Tanenhaus MK, Aslin RN. Tracking the timecourse of phonetic cue integration during spoken word recognition. Psychonomic Bulletin and Review. 2008; 15(6): 1064–1071. [PubMed: 19001568]

McMurray B, Samelson VS, Lee SH, Tomblin JB. Eye-movements reveal the time-course of online spoken word recognition language impaired and normal adolescents. Cognitive Psychology. 2010; 60(1):1–39. [PubMed: 19836014]

McMurray B, Tanenhaus MK, Aslin RN. Gradient effects of within-category phonetic variation on lexical access. Cognition. 2002; 86(2):B33–B42. [PubMed: 12435537]

McMurray B, Tanenhaus MK, Aslin RN. Within-category VOT affects recovery from “lexical” garden paths: Evidence against phoneme-level inhibition. Journal of Memory and Language. 2009; 60(1): 65–91. [PubMed: 20046217]

Mirman D, Dixon J, Magnuson JS. Statistical and computational models of the visual world paradigm: Growth curves and individual differences. Journal of Memory and Language. 2008; 59(4):475–494. [PubMed: 19060958]

Nation K, Marshall CM, Altmann GTM. Investigating individual differences in children’s real-time sentence comprehension using language-mediated eye movements. Journal of Experimental Child Psychology. 2003; 86(4):314–329. [PubMed: 14623215]

Neely JH. Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. Journal of Experimental Psychology: General. 1977; 107(3):226–254.

Praamstra P, Meyer AS, Levelt WJM. Neurophysiological manifestations of phonological processing: Latency variation of a negative ERP component time locked to phonological mismatch. Journal of Cognitive Neuroscience. 1994; 6:204–219. [PubMed: 23964972]

Revill KP, Spieler DH. The effect of lexical frequency on spoken word recognition in young and older listeners. Psychology and Aging. 2012; 27(1):80–87. [PubMed: 21707175]

Robertson EK, Joanisse MF, Desroches AS, Ng S. Categorical speech perception deficits distinguish language and reading impairments in children. Developmental Science. 2009; 12(5):753–767. [PubMed: 19702768]

Salverda AP, Dahan D, McQueen JM. The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. Cognition. 2003; 90:51–89. [PubMed: 14597270]

Sedivy JC, Tanenhaus MK, Chambers CG, Carlson GN. Achieving incremental semantic interpretation through contextual representation. Cognition. 1999; 71(2):109–147.10.1016/s0010-0277(99)00025-6 [PubMed: 10444906]

Semel, E.; Wiig, E.; Secord, WA. Clinical Evaluation of Language Fundamentals 4 (CELF-4). San Antonio, TX: The Psychological Corporation; 2003.

Smith LB, Quittner AL, Osberger MJ, Miyamoto R. Audition and visual attention: The developmental trajectory in deaf and hearing populations. Developmental Psychology. 1998; 34(5):840–850. [PubMed: 9779732]

Spivey-Knowlton, MJ.; Allopenna, P. A computational account of the integration of linguistic and visual information during spoken word recognition. Proceedings of the Computational Psycholinguistics Conference; Berkeley, CA. 1997.

Spivey, MJ. The Continuity of Mind. New York: Oxford University Press; 2007.

Spivey MJ, Grosjean M, Knoblich G. Continuous attraction toward phonological competitors. Proceedings of the National Academy of Sciences of the United States of America. 2005; 102(29): 10393–10398.10.1073/pnas.0503903102 [PubMed: 15985550]

Tanenhaus MK, Spivey-Knowlton MJ, Eberhard KM, Sedivy JC. Integration of visual and linguistic information in spoken language comprehension. Science. 1995; 268:1632–1634. [PubMed: 7777863]

Tanenhaus, MK.; Trueswell, JC. Sentence Comprehension. In: Eimas, PD.; Miller, JL., editors. Handbook in perception and Cognition. Speech, language and communication. Vol. 11. New York, NY: Academic Press; 1995. p. 217-262.

Toscano JC, McMurray B. Online integration of acoustic cues to voicing: Natural vs. synthetic speech. Attention, Perception & Psychophysics. (in press).

Trueswell JC, Sekerina I, Hill NM, Logrip ML. The kindergarten-path effect: Studying on-line sentence processing in young children. Cognition. 1999; 73:89–134. [PubMed: 10580160]

Ullman MT, Pierpont EI. Specific language impairment is not specific to language: The procedural deficit hypothesis. Cortex. 2005; 41:399–433. [PubMed: 15871604]

Van Petten C, Kutas M. Interactions between sentence context and word frequency in event-related brain potentials. Memory and Cognition. 1990; 18:380–393. [PubMed: 2381317]

Vanderwart M. Priming by pictures in lexical decision. Journal of Verbal Learning and Verbal Behavior. 1984; 23:67–83.

Yee E, Blumstein S, Sedivy J. Lexical-semantic activation in Broca's and Wernicke's aphasia: Evidence from eye movements. Journal of Cognitive Neuroscience. 2008; 20(4):592–612. [PubMed: 18052783]

Yee E, Sedivy J. Eye movements to pictures reveal transient semantic activation during spoken word recognition. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2006; 32:1–14.

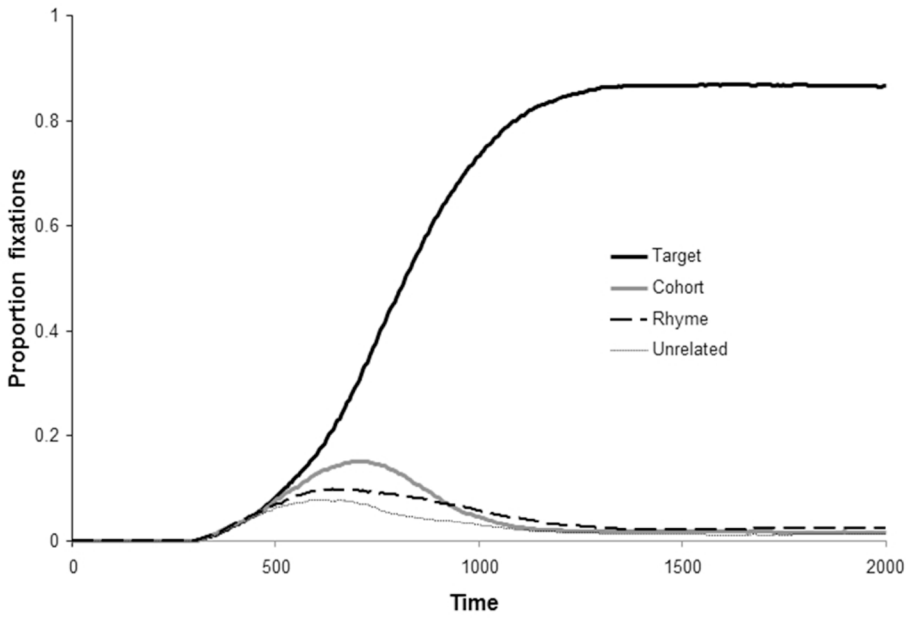

Figure 1.
Fixations to target, cohort, rhyme, and unrelated objects in the auditory paradigm, averaged across days.

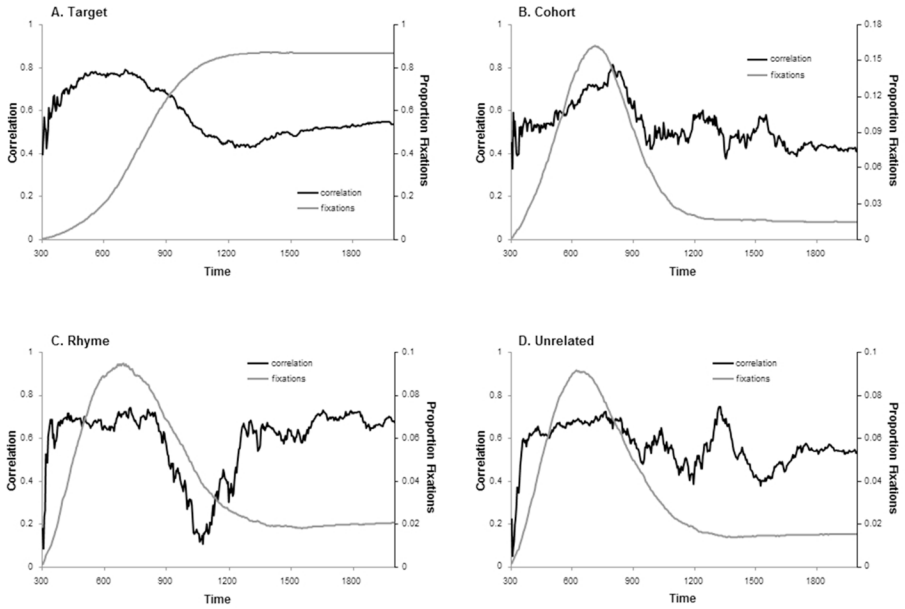

Figure 2.
Time-slice correlations overlaid on looks to the target (A), cohort (B), rhyme (C) and unrelated (D) items.

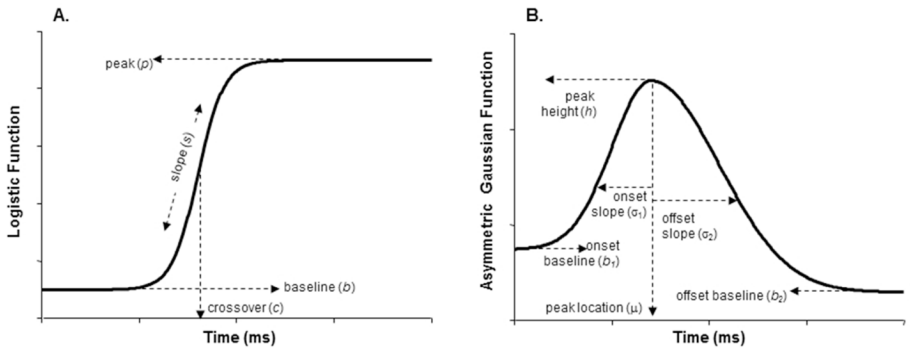

Figure 3.
Logistic (A) and asymmetric Gaussian (B) functions for curvefitting.

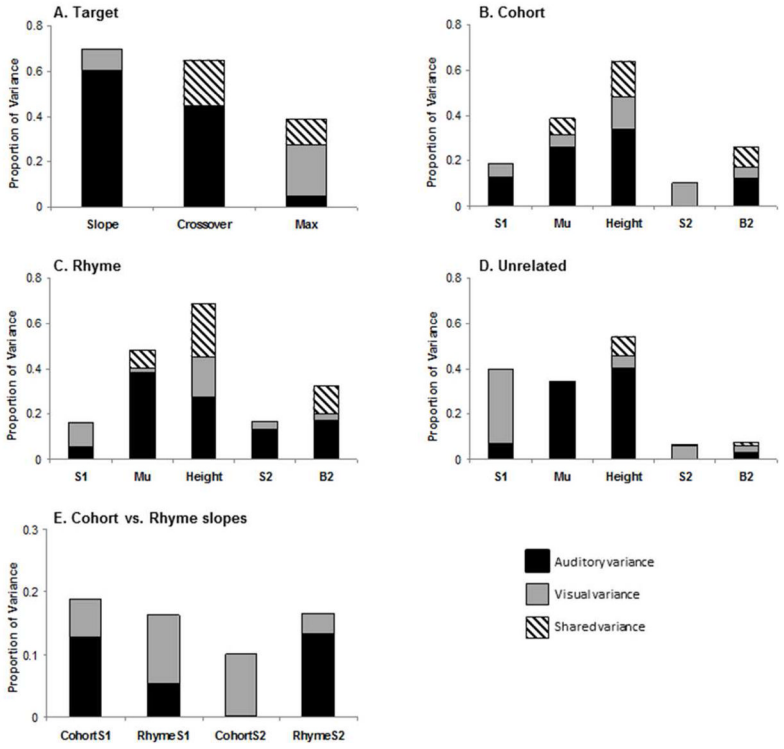

Figure 4.
Proportion of variance accounted for by auditory and visual parameters and additional shared variance.

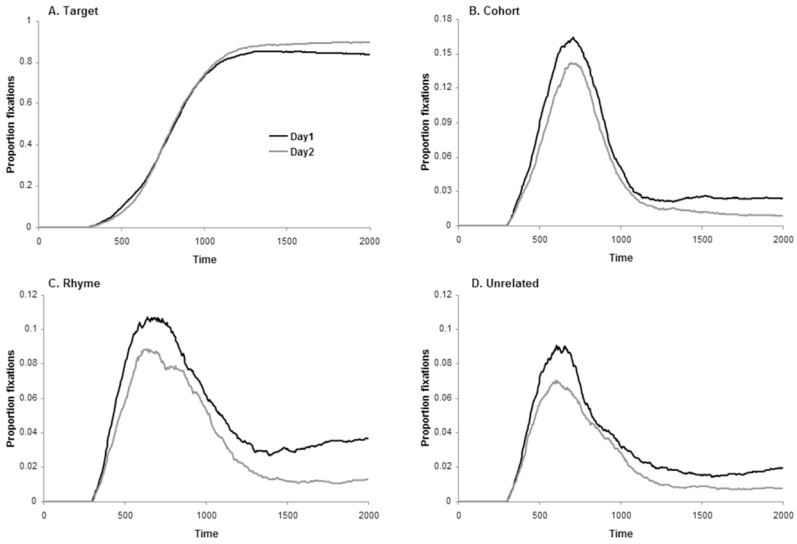

Figure 5.
Fixations to target (A), cohort (B), rhyme (C) and unrelated (D) items on Day 1 and Day 2.)

Table 1

Evidence for Lexical Activation in Both Non-eye-tracking and Eye-tracking Paradigms. Shown are representative papers for each finding, and not a complete bibliography.

Phenomenon	Non-eye-tracking	Eye-tracking
cohort activation	Marslen-Wilson & Zwitserlood (1989); Praamstra, Meyer & Levelt (1994); Desroches, Newman & Joanisse (2008);	Allopenna et al. (1998); McMurray et al. (2010)
rhyme activation	Marslen-Wilson, Moss & Van Halen (1996); Connine, Blasko & Titone (1993); Praamstra, Meyer & Levelt (1994); Desroches, Newman & Joanisse (2008);	Allopenna et al. (1998); McMurray et al. (2010)
neighborhood density effects	Luce & Pisoni (1998)	Magnuson et al. (2007); Apfelbaum, Blumstein & McMurray (2011)
frequency effects	Marslen-Wilson (1987); Van Petten & Kutas (1990)	Dahan, Magnuson & Tanenhaus, (2001); Gaskell & Dahan (2007)
sensitivity to sub- phonetic detail	Andruski, Blumstein & Burton (1994);	McMurray et al. (2008; 2002)
embedded word effects	Gow & Gordon (1995)	Salverda, Dahan & McQueen, (2003)
semantic priming	Neely (1977); Vanderwart (1984); Connolly & Phillips (1994); Kutas & Hillyard (1984)	Apfelbaum, Blumstein & McMurray (2011); Yee & Sedivy (2006)

Table 2

Role of Words and Shapes by Trial Type

a. Auditory/Lexical Task

			Visual stimuli			
		Auditory stimulus	**horn**	**horse**	**corn**	**box**
Trial Type	**TCRU**	**horn**	target	cohort	rhyme	unrelated
	TCUU	**horse**	cohort	target	unrelated	unrelated
	TRUU	**corn**	rhyme	unrelated	target	unrelated
	TUUU	**box**	unrelated	unrelated	unrelated	target

b. Visual Task

			Visual stimuli			
		Flashed stimulus	**red triangle**	**red circle**	**blue triangle**	**yellow square**
TrialType	**TClrSU**	**red triangle**	target	color match	shape match	unrelated
	TClrUU	**red circle**	color match	target	unrelated	unrelated
	TSUU	**blue triangle**	shape match	unrelated	target	unrelated
	TUUU	**yellow square**	unrelated	unrelated	unrelated	target

Note. T=target, C=cohort, R=rhyme, U=unrelated

Note. T=target, Clr=color, S=shape, U=unrelated

J Speech Lang Hear Res. Author manuscript; available in PMC 2014 August 01.

Table 3

Target and Competitor Parameter Correlations Across the Two Auditory Tasks

	Parameter	R	p
Target	slope (Δp[fix] / ms)	.751	<.001
	crossover (ms)	.826	<.001
	asymptote (p[fix])	.390	.037
Cohort	onset slope (ms)[a]	.339	.072
	midpoint (ms)	.616	<.001
	peak height (p[fix])	.703	<.001
	offset slope (ms)	−.057	.768
	offset baseline (p[fix])	.483	.008
Rhyme	onset slope	.206	.283
	midpoint	.702	<.001
	peak height	.716	<.001
	offset slope	.346	.066
	offset baseline	.554	.002
Unrelated	onset slope	.110	.572
	midpoint	.562	.002
	peak height	.716	<.001
	offset slope	.075	.699
	offset baseline	.169	.381

[a]The competitor onset and offset slope measurements correspond to the standard deviation (σ) or width of a Gaussian and are therefore represented in ms.

Table 4

Results of Regression Analysis on Auditory and Visual Tasks

	Step:	*1*	*2*	*3*	Total R^2	Raw Reliability	Updated Reliability
	Parameter	**V Day (df=1,27)**	**A1&V**	**V Day x A1**			
Target	slope	.010	.675*	.058	.743	.751	.862
	crossover	.052	.695*	.002	.749	.826	.865
	asymptote	.039	.384*	.037	.460	.390	.678
Cohort	onset slope	.001	.174	.013	.188	.339	.434
	midpoint	.068	.387	.014	.469	.616	.685
	height	.075	.639*	<.001	.714	.703	.845
	offset slope	.028	.100	.106†	.235	−.057	.485
	offset baseline	.024	.260†	.083†	.366	.483	.605
Rhyme	onset slope	.155*	.145	.004	.304	.206	.552
	midpoint	.063	.480*	.008	.551	.702	.742
	height	.033	.685*	.007	.725	.716	.851
	offset slope	.009	.160	.137*	.306	.346	.553
	offset baseline	.010	.325*	.049	.384	.554	.620
Unrelated	onset slope	.098†	.349*	.043	.490	.110	.700
	midpoint	.066	.343*	.116*	.525	.562	.725
	height	.042	.540*	.002	.584	.716	.764
	offset slope	.002	.065	.147*	.215	.075	.463
	offset baseline	.036	.073	.001	.110	.169	.332

Note. For each auditory Day 2 parameter, R^2_Δ is shown after each variable was entered into the model. Step 2 target: df=2,25; competitors: df= 3,24. Step 3 target: df=1,24; competitors: 1,23. Total R^2 is in the antepenultimate column, followed by the R values calculated in the prior analysis and finally the updated R values obtained in the regression analysis.

Significance of the F-statistic on change:

* $p < .05$, two-tailed;

† p < .1 (marginal), two tailed.

Table 5

Uniquely Auditory, Uniquely Visual, and Shared Variance

	Parameter	Total	Unique Auditory	Unique Visual	Shared
Target	slope^{o}	.675*	.604*	.093*	--
	crossover	.695*	.612*	.054*	.029
	asymptote	.384*	.050	.226*	.108
Cohort	onset slope^{o}	.174	.128†	.060	--
	midpoint	.387	.264*	.049	.074
	height	.639*	.340*	.141*	.158
	offset slope^{o}	.100	.002	.099	--
	offset baseline	.260†	.122†	.050	.088
Rhyme	onset slope^{o}	.145	.053	.11	--
	midpoint	.480*	.383*	.018	.080
	height	.685*	.275*	.176*	.233
	offset slope^{o}	.160	.133†	.034	--
	offset baseline	.325*	.172*	.029	.124
Unrelated	onset slope^{o}	.349*	.070†	.328*	--
	midpoint	.343*	.336*	.007	0
	height	.540*	.401*	.056†	.083
	offset slope	.065	.001	.062	.002
	offset baseline	.073	.027	.036	.01

Note. Significance of the F-statistic on change:

* $p < .05$, two-tailed;

† $p < .1$ (marginal), two tailed.

^{o} indicates negative correlations between A1 and V or between A2 and V.

J Speech Lang Hear Res. Author manuscript; available in PMC 2014 August 01.

Table 6

Results of t-Tests Comparing Auditory Day 1 and 2 Parameters

		Mean		t(28)	p
	Parameter	**Day 1**	**Day 2**		
Target	slope	.0019	.0021	−3.9	.001
	crossover	769	770	−.268	.791
	asymptote	.858	.879	−.888	.382
Cohort	onset slope	184	164	1.625	.115
	midpoint	708	728	−1.446	.159
	peak height	.1826	.1610	2.315	.028
	offset slope	184	164	1.868	.072
	offset baseline	.0175	.0128	1.917	.066
Rhyme	onset slope	162	134	2.201	.036
	midpoint	718	711	.271	.789
	peak height	.1218	.0987	3.748	.001
	offset slope	204	219	−.867	.393
	offset baseline	.0263	.0120	5.084	<.001
Unrelated	onset slope	146	130	1.038	.308
	midpoint	676	685	−.286	.777
	peak height	.1105	.0922	2.807	.009
	offset slope	231	202	1.789	.084
	offset baseline	.0169	.0093	3.051	.005

J Speech Lang Hear Res. Author manuscript; available in PMC 2014 August 01.