CHAPTER 4

Looking while listening

Using eye movements to monitor spoken language comprehension by infants and young children

Anne Fernald, Renate Zangl, Ana Luz Portillo
and Virginia A. Marchman

The "looking-while-listening" methodology uses real-time measures of the time course of young children's gaze patterns in response to speech. This procedure is low in task demands and does not require automated eyetracking technology, similar to "preferential-looking" procedures. However, the looking-while-listening methodology differs critically from preferential-looking procedures in the methods used for data reduction and analysis, yielding high-resolution measures of speech processing from moment to moment, rather than relying on summary measures of looking preference. Because children's gaze patterns are time-locked to speech and coded frame-by-frame, each 5-min experiment response latencies can be coded with millisecond precision on multiple trials over multiple items, based on data from thousands of frames in each experiment.

The meticulous procedures required in the collection, reduction, and multiple levels of analysis of such detailed data are demanding, but well worth the effort, revealing a dynamic and nuanced picture of young children's developing skill in finding meaning in spoken language.

1. Introduction

Developmental studies of comprehension in very young children have relied traditionally on off-line measures, responses made after the offset of the speech stimulus that do not tap into the real-time properties of spoken language. Studies of incremental processing by adults rely on on-line measures that monitor the time course of the listener's response in relation to key points in the speech signal. Because comprehension occurs rapidly and automatically without time for reflection, it is revealing to study the listener's interpretation during speech processing...
and not just afterward. Classic on-line behavioral techniques used to investigate incremental speech processing by adults include phoneme monitoring (Cutler & Foss 1977), gating (Grosjean 1983), and cross-modal priming (Marslen-Wilson & Zwitserlood 1989), among others. Although some of these have been adapted for use with school-aged children (Cutler & Winney 1987; Wulley 1993; Chaisen this volume), the task demands are often problematic for younger children. Research with adults using automated eyetracking techniques has been extremely productive in recent years, providing sensitive on-line measures high in ecological validity (Tanenhaus, Magnuson, Dahan & Chambers 2000). This technique has also been with 4–5-year-old children (Trueswell, Sekerina, Hill & Logrip 1999; Snedeker & Trueswell 2004).

Here we describe a simpler but equally powerful experimental method for monitoring the time course of comprehension by infants and very young children, which we call the looking-while-listening procedure. In this procedure, children look at pairs of pictures while listening to speech naming one of the pictures. Their gaze patterns are videotaped and eye movements are measured with high precision in relation to relevant points in the speech signal. Using the looking-while-listening paradigm, we have shown that speed and efficiency in infants’ on-line responses to familiar words in continuous speech increase dramatically over the second year for both English- and Spanish-learning children (Fernald, Pinto, Swingley, Weinberg, & McRoberts 1998; Hurtado, Marchman, & Fernald 2007), that 24-month-olds are able to process phonetic information incrementally (Swingley, Pinto, & Fernald 1999; Fernald, Swingley, & Pinto 2001), and that individual differences in children’ speech processing efficiency are related to their level of lexical and grammatical development (Fernald, Perfors, & Marchman 2006). These and other recent findings validate the looking-while-listening paradigm as a powerful new method for exploring how young children take advantage of perceptual and linguistic features of the speech they hear in relation to information in the visual world, as they learn to find meaning in spoken language.

1.1 Research on the early development of receptive language skills

Long before they can speak their first words, infants begin to reveal their emerging knowledge of language by responding meaningfully to the speech they hear. Yet because comprehension is a mental event that can only be inferred indirectly from a child’s behavior in a particular context, the early development of receptive language competence has been less accessible to scientific inquiry than developmental gains in speech production. However, research on infant cognition over the past three decades has led to valuable experimental methods for “reading the minds” of very young language learners. Techniques that have made it possible to explore the developmental origins of understanding in greater depth and with greater precision. Much of this research has examined how infants become attuned to sound patterns in the ambient language over the first year of life (e.g., Werker & Tees 1984; Kuhl 2004), and attend to speech patterns relevant to language structure (e.g., Jusczyk 1997; Saffran 2002). These studies show that over the first year infants become skilled listeners, able to make distributional analyses of phonetic features of spoken language, and that they form some kind of acoustic-phonetic representation for frequently heard sound patterns (e.g., Halle & de Boysson-Bardies 1994). Such accomplishments are often cited as evidence for early “word recognition”. But since this selective response to familiar words can occur without any association between particular sound patterns and meanings, it is perhaps appropriately viewed as evidence for pattern-detection abilities prerequisite for recognizing words in continuous speech. Much less is known about how language-specific processing strategies continue to develop beyond the first year, as children in their second and third years begin to appreciate regularities at higher levels of linguistic organization, using their emerging lexical and morphosyntactic knowledge to make sense of spoken words and sequences of words in combination (see Fernald & Frank in press: Naglies 2002).

1.2 Observational and off-line experimental measures of children’s skill in comprehension

Between the ages of 10 and 14 months children typically begin to show signs of associating sound patterns with meanings, speaking a few words and appearing to understand many more. By the end of the second year, they reveal progress in understanding through increasingly differentiated verbal and behavioral responses to speech. But growth in receptive language competence is harder to observe than growth in productive abilities, because the processes involved in comprehension are only partially and inconsistently apparent through the child’s spontaneous behavior. Scientific studies of early comprehension have made use of quite different methodologies that fall into four main categories (see Fernald 2002): Diary studies provided the first systematic observational data on early comprehension abilities, describing how young children appear to interpret speech in their everyday activities and interactions (e.g., Lewis 1936; Bloom 1973). Studies of vocabulary growth use parental-report checklists to track changes in the estimated size of the child’s receptive lexicon; such changes correlate in interesting ways with later grammatical development (e.g., Fenson, Marchman, Thal, Dale, Reznick, & Bates 2007). Naturalistic experiments on comprehension use behavioral responses to test
infants’ ability to identify familiar words (e.g., Benedict 1979) and understand words in combination (e.g., Shipley, Smith, & Gleitman 1969). Experiments on word learning constitute the largest area of research related to early understanding: one approach focuses on how cognitive biases guide inferences about word meanings (e.g., Markman 1989); others explore how children use linguistic (e.g., Katz, Baker, & McNamara 1974) and pragmatic knowledge (e.g., Tomasello 2000) to guide these inferences.

All of these approaches rely on off-line measures of comprehension, i.e., assessments of understanding based on observing the child’s behavior after hearing a particular linguistic stimulus. In the case of diary observations and parental-report checklists, the judgment that a child does or does not “understand” a word such as dog or cup is made informally by adults who interact regularly with the child in many different contexts. In the case of off-line measures used experimentally, these judgments are based on the child’s behavior in a more controlled situation, with a clearly defined response measure such as choosing an object or pointing to a picture given two or more alternatives. What these measures have in common is that they all are based on children’s responses to a spoken word or sentence after it is complete rather than as it is heard and processed. While such off-line procedures enable researchers to assess whether or not a child responds systematically in a way that indicates understanding, they do not tap into the real-time properties of spoken language and thus reveal less about the child’s developing efficiency in identifying and interpreting familiar words in continuous speech.

The development of preferential-looking measures for assessing comprehension

In 1963 Robert Fantz published the first study using a preferential-looking method with infants, showing that even newborns looked selectively at some visual stimuli over others. Although Haith (1980) later questioned whether “preference” was the appropriate way to characterize infants’ selective looking behavior, the findings that emerged from dozens of preferential-looking studies in this period suggested that certain early visual biases appeared to be independent of previous experience with particular stimuli. Adapting the preferential-looking procedure to investigate cross-modal perception in infants, Spelke (1976) presented infants with two visual stimuli, only one of which matched a simultaneously presented auditory stimulus, and found that infants looked significantly longer to the matching than to the non-matching visual stimulus. This auditory-visual matching procedure was later modified in different ways by investigators interested in the development of language comprehension in the early years of life.

The first experimental procedures for testing infants’ knowledge of object words in a controlled setting were also introduced at the end of the 1970s. Benedict (1979) found that 12-month-olds would orient reliably to a familiar object when it was named, even when nonverbal behaviors such as gaze and pointing by the speaker were eliminated. In a more precisely controlled procedure using eye movements as an index of word recognition, Thomas, Campos, Shucard, Ramsey and Shucard (1981) compared the ability of 11- and 13-month-old infants to identify a familiar named object from an array of competitors matched for visual salience. The finding that some 12-month-olds could identify the correct referent of a few familiar words was hardly surprising, since that much was known from observational studies and parental report. What this new method offered was a way to assess word recognition more objectively. Unlike the informal observations of comprehension used earlier, this procedure enabled the experimenter to standardize stimulus presentations, to define carefully which behaviors counted as a correct response, and to eliminate gestural and other nonverbal cues from the experimenter that might indicate the target object.

The innovative studies by Spelke (1976) and Thomas et al. (1981) provided a foundation for later research by Golinkoff, Hirsh-Pasek, Cauley and Gordon (1987) and by Reznick (1990) in which preferential-looking measures were further adapted for use in assessing early language comprehension. The version of the method developed by Golinkoff et al., now known as the “intermodal preferential looking paradigm” (IPLP), has been used in numerous studies (e.g., Hollich, Hirsh-Pasek, & Golinkoff 2000; Shafer & Plunkett 1998; Meints, Plunkett, & Harris 2002). In the procedures used in these studies, infants are shown two pictures of objects as they hear speech naming one of the objects. In some studies gaze patterns are coded in real-time using a button box to record fixations on the target and distractor objects and shifts between the two pictures. The dependent measures typically used as an index of comprehension in such studies include total looking time to the target picture and duration of longest look to the target picture. With these measures, researchers using the IPLP have been able to investigate several interesting questions about children’s early lexical and syntactic knowledge – for example, whether or not infants at a particular age look relatively longer to the correct target picture than to the distracter picture when asked to identify the target object in a sentence such as Find the cat (e.g., Golinkoff et al. 1987), or to interpret a spatial proposition in a sentence such as Look at the cat on the table (Meints et al. 2002).
1.4 The evolution of the looking-while-listening procedure: Moving to real-time measures

Informed by the studies of Thomas et al. (1981), Golinkoff et al. (1987), and Reznick (1990), our research group began to use a modified version of the preferential-looking method to investigate whether particular features of infant-directed speech, such as exaggerated pitch and vowel lengthening, might make it easier for young language learners to identify familiar words in fluent speech. Our initial goal in modifying the preferential-looking paradigm was to increase the sensitivity, reliability, and validity of the measures, by making minor modifications to the procedure that served to eliminate confounding variables. Earlier preferential-looking studies had used different stimuli as target and distracter objects, thus potentially confounding object salience with target status. And some had also failed to counterbalance side of target object presentation, another confound that made it difficult to interpret infants' selective looking behavior unambiguously. To address these concerns, we made sure that all target objects were also presented as distracters to reduce the influence of initial object preferences, and that side of presentation was fully counterbalanced to control for possible side bias. These and other minor but potentially influential changes were undertaken to increase internal validity in our experimental designs. In addition, we made a major change in the measures used to capture infants' gaze patterns in response to speech: Rather than coding eye movements in real-time using a button box, as was the practice in other preferential-looking procedures at the time, we began to code eye movements from the videotapes, frame-by-frame in slow motion. This change enabled us to eliminate from our measurements the noise introduced by the ca. 300 ms latency of the observer to press the button, a critical first step in the direction of achieving greater precision in our dependent measures. These modifications resulted in a labor-intensive version of the original procedure, requiring several hours to code the record of each 5-min test session, but the enhanced reliability of the measurements justified the effort.

However, we were still not yet taking full advantage of the potential for increased temporal resolution in our measures of infants' gaze patterns. In the Golinkoff et al. (1987) paradigm, word recognition was operationalized as a tendency to look longer at the named target picture than at the distractor picture, with looking-time-to-target averaged over a 6-s measurement window following offset of the speech stimulus. A similar criterion is common today in most labs using this procedure (Hoffich et al. 2000; Schaer & Plunkett 1998), and we adopted it initially as well. The 6-s response window reflected a premise that seemed entirely reasonable at the time. Although psycholinguistic research with adults had recently shown that experienced listeners can process speech incrementally, generating hypotheses about word identity based on what they have heard up to that moment (Marlens-Wilson & Zwitserlood 1989), we simply assumed that infants would be considerably slower than adults in processing continuous speech. And since there were no data yet available on the time course of real-time speech processing by infants, we had to discover the hard way, through a long process of trial and error, how wrong we were in this assumption!

In three early studies we tested infants at different ages to investigate the influence of prosodic features on their ability to recognize familiar words. While these results were promising (Fernald, McRoberts, & Herrera 1992; see Fernald, McRoberts, & Swingley 2001, for a review of these findings), it became increasingly clear that we needed to understand better how the time course of infants' responses changed with development over the second year. For example, when using a percent-correct measure averaged over a 6-s measurement window, we were surprised to find that 24-month-olds apparently performed less well than 18-month-olds. This counterintuitive result suggested that our measures of looking time were failing to capture the gains in accuracy expected toward the end of the second year, an observation that led us to make parametric reductions in the time window over which looking-time was averaged. When we used a 4-s measurement window, the data began to make more sense, and with a 2-s window, the predicted improvement in word recognition finally became clear: for 18-month-olds the mean percentage of looking time to the target picture during this much shorter window was ca. 60%, and for 24-month-olds it rose to 80%. By adopting the 6-s coding window that was standard at the time, we had greatly underestimated the accuracy of 24-month-olds in this word recognition task; these older infants had in fact oriented quickly to the target picture upon hearing it named and had looked at it for 2–3 seconds, but then they tended to look back at the other picture or to look away. Since these look-backs and look-aways by the 24-month-olds in most cases followed a correct response within a few seconds, they were actually a sign of rapid processing; however, when this post-response "noise" was averaged into the percent-looking-to-target averaged over a 6-s window, the 24-month-olds looked less overall and thus appeared to be less accurate than 18-month-olds. This discrepancy was obviously not because the older children were slower or less reliable in recognizing familiar words, but rather because they responded more quickly than the younger children and then tended to lose interest in the target picture midway through the 6-s window.

It was clear from these analyses that infants were able to identify familiar spoken words much more rapidly than we had imagined possible. Thus at this point we made two important procedural changes: First, we began to measure infants' eye movements from the onset of the target word, not the offset. And second, rather than averaging looking time over an arbitrary, fixed coding window (which
might be suitable at one age but not at another), we began to code eye movements at the finest level of resolution possible given the limits of our technology. In the first published study using time course measures of spoken word recognition with infants (Fernald et al. 1998), this limit was 100 ms, the resolution of our time code generator at that time. In subsequent studies the level of resolution increased to 33 ms, the duration of a single video frame. This change was critical in enabling us to move from the global measure of total looking time to the target picture to a much more precise measure of reaction time, capturing the child’s latency to shift from the distracter to the target picture. Through this incremental process of refining our analysis techniques and improving our coding technology, the looking-while-listening procedure has become an increasingly powerful method for monitoring the time course of infants’ comprehension of continuous speech, enabling us to measure reaction time as well as accuracy in word recognition in very young language learners. The result is a testing procedure with minimal task demands that can be used with both infants and adults, yielding eye movement data comparable in reliability and precision to data from adult studies that require technically more sophisticated automated eyetracking methods (e.g., Tanenhaus et al. 2000; Henderson & Ferreira 2004).

2. From preference to reference: A nuts-and-bolts overview of the looking-while-listening paradigm

In this section we provide an overview of the looking-while-listening procedure, following the traditional format of a Methods section in a research report (i.e., describing participants, procedure, etc.). However, consistent with the focus on methodology in this volume, the information in each section extends beyond the details relevant to any particular study, integrating our experience using this paradigm with many different experimental designs and with participants ranging from 12-month-old infants to adults. Our goal here is to present an overview of the looking-while-listening paradigm at a functional level, discussing the logic of each step in the procedure, from preparing and running an experiment to coding eye movements and analyzing the data using several different measures of efficiency in spoken language processing.

As mentioned in the introduction, the looking-while-listening procedure is superficially similar to a preferential-looking procedure in that infants are shown two pictures on each trial and hear speech naming one of the pictures as we record their gaze patterns in response to the speech signal. But the static notion of “preference” is irrelevant for our purposes. Rather than constraining infants’ looking behavior in response to spoken language as motivated by preference, we are interested in how children establish reference by making sense of spoken language from moment to moment, a process of incremental interpretation that is highly dynamic in adult comprehension (e.g., Knoeberle & Crocker 2006). Rather than relying exclusively on a single preference score based on total looking to the target averaged over a fixed coding window following the speech stimulus, we are also interested in the time course of looking to the referent as the sentence unfolds. To achieve this goal we code children’s gaze patterns off-line through careful frame-by-frame inspection, enabling precise measurement of their latency to initiate an eye movement toward the appropriate referent, time-locked to critical moments in the speech signal on each trial. Thus the looking-while-listening procedure incorporates the same sensitive temporal measures used in eyetracking studies with adults and older children (see Trueswell & Tanenhaus 2005), differing from those methods in only three noteworthy respects: first, we typically use visual displays with only two alternatives rather than more complex scenes involving four or more displays; second, we do not use an automated eyetracker, because comparable precision and reliability can be achieved using our high-resolution video coding procedures; and third, an inherent constraint in working with infants and very young children is that they have limited and fluctuating attention; thus we have to design experiments lasting just a few minutes with only 30–40 trials, which yield much less (and potentially noisier) data than the longer experiments with hundreds of trials that are possible in studies with adults.

2.1 Participants

The looking-while-listening paradigm has been used effectively in our laboratory with infants as young as 14 months of age as well as older children and adults. When we tested 12-month-olds in several pilot studies using this procedure, we found their performance to be close to chance, as did Zangl, Klaman, Thal, Fernald and Bates (2005) in a study using a similar procedure. While 12-month-olds are happy to look at the pictures displayed, they are likely to fixate on only one picture and ignore the other on any given trial, shifting less frequently overall than infants just two months older. This pattern of results with younger infants reflects their limited linguistic knowledge, but may also result from attentional limitations at this age.

2.2 Preparing visual stimuli

In studies in which we assess recognition of familiar words, the visual stimuli consist of pictures of real objects with which infants and toddlers are very famil-
iar. The target objects chosen for each study have names that are highly likely to be understood by children in the relevant age range, based on vocabulary norms for the MacArthur-Bates Communicative Development Inventory (Fenson et al. 2007). Realistic images of common objects judged to be prototypical for children at each age are drawn from various sources, including image banks on the Internet and pictures taken with digital cameras. For studies in which children are taught and then tested on novel words, we use pictures of constructed objects with which children have no prior experience. In some cases, objects are shown against a uniform gray background; however, in studies where the same target objects are used repeatedly across trials, objects may be presented on more diverse and complex backgrounds that vary in color and pattern, in order to maintain children’s interest.

All images used as stimuli are edited so the objects are approximately the same size and are informally matched in visual complexity and brightness. When choosing and editing pictures for a particular study, it is important to balance the visual salience of the target and distracter pictures on each trial, keeping in mind that any given picture is not inherently “salient” but only in relation to other pictures in the stimulus set. If one picture in a pair is much more engaging than another, it is more difficult to tell whether looking to the target picture on a given trial was influenced by the speech stimulus rather than by baseline visual preferences. We have found that younger infants in particular generally find images of animate objects more interesting than inanimate ones, and thus might initially attend more to a picture of an animal if it was paired with a picture of an artifact. To reduce the effects of this potential bias, we present yoked pairs of animates (dog and baby) on some trials, and pairs of inanimates (shoe and car) on others. Although counterbalancing measures, such as presenting each stimulus picture as both target and distracter an equal number of times in every experiment, are designed to mitigate the effects of differential visual salience among stimuli, such differences should be reduced as much as possible because they contribute noise to the data.

In some experiments the potential for visual salience differences is particularly acute due to the nature of the stimulus objects used, and in this case pilot testing is advisable. To evaluate the suitability of new images as potential stimuli, candidate pictures are paired with other potential stimulus pictures and presented to children in the appropriate age range without any verbal information, to determine whether participants orient more to one picture than the other based on differences in visual salience. If children do indeed fixate one picture significantly more than the other, the pictures need to be edited further to adjust factors like hue, saturation, and visual complexity that might contribute to differential salience. The modified stimuli then need to be pilot-tested again until children are equally likely to fixate both pictures. Balancing the relative visual salience of stimulus objects is an important step in maximizing the sensitivity of the measures in the looking-while-listening procedure, especially when working with infants younger than two years of age. This extra step is not as labor-intensive as it may seem, if such “picture-pilot” trials for potential stimuli in a new experiment can be included as filler trials in another ongoing experiment with children in the same age range.

2.3 Preparing auditory stimuli

In the looking-while-listening procedure children’s eye movements are examined in relation to words that are heard at critical points in the unfolding speech stream, and so precision in recording and editing the speech stimuli is key for all further analyses. To produce the stimuli used in our studies, multiple tokens of each sentence are digitally recorded by a native speaker of English or Spanish, then acoustically analyzed, and edited using Peak or Praat sound-editing software. Because it is a goal in most of our studies to achieve comparability in relevant acoustic dimensions across the set of speech stimuli, recording ten or more tokens of each stimulus sentence yields a range of candidate stimuli to choose from. After measuring each relevant element in every recording, we then choose the tokens of each sentence type that are most closely matched in the duration of critical words as well as in overall prosodic characteristics, across the whole stimulus set. The criteria for final stimulus selection vary to some extent with the experimental question. For example, in a recent study (Zangl & Fernald 2007) investigating children’s responses to grammatical and non–determiners (Where’s the dog? vs. Where’s po’ dog?), the choice of tokens for the final stimulus set was governed by three considerations: First, all carrier frames were matched in duration ($M = 269$ ms; range: $261$ to $278$). Second, grammatical and non–determiners were matched in duration (grammatical: $M = 160$ ms; range: $147$ to $172$; non–determiner: $M = 168$ ms; range: $156$ to $173$). Third, there was also important that all articles be unstroressed, and to the same degree. When rated by two phonetically trained listeners, both sentence types were judged to be comparable in intonation contour. In some studies, cross-splicing techniques are also used to increase stimulus control. In the Zangl and Fernald study, for example, it was important to ensure that each target noun was acoustically identical for each sentence type, and so a single token of each target noun was cross-spliced into carrier frames across conditions.

Depending on the goals and design of the study, measurements are made for particular regions of interest in each stimulus sentence, such as carrier frames, determiners, adjectives, target nouns, etc., to enable reliable identification of the
acoustic onset and offsets of key words that are crucial for measures of response latency. For example, when investigating children’s efficiency in interpreting familiar object names in simple English sentences such as *Find the car*, determining the onset and duration of the target word is sufficient to time-lock the child’s response to the unfolding object name (e.g., Fernald et al. 1998, 2006). In this case, the noun onset is the critical point at which the child begins to accumulate phonetic information necessary to interpret the target word and identify the referent object. However, the critical point may come earlier in cases where linguistic information encountered before the noun is potentially informative. For example, nouns in Spanish have obligatory grammatical gender and are preceded by gender-marked determiners. In a study investigating whether Spanish-learning children could take advantage of gender-marked articles to identify an upcoming object name, Lew-Williams and Fernald (2007) measured children’s eye movements to the target object in relation to the first potentially informative element in the sentence, in this case the prenominal article. This study required measurement of the onset and offset of the determiner as well as the target noun, and tokens of each stimulus sentence were chosen to maximize comparability on these dimensions across the stimulus set. That is, the best-matching tokens were selected from the larger set of exemplars recorded for each sentence. These were then further edited so that key points in the speech stream could be carefully matched within one frame across sentences.

By making the critical speech stimuli in each condition prosodically comparable, we can reduce the effects of potentially confounding factors that might influence children’s responses. However, maximizing experimental control in this way also has a less desirable consequence: we end up with a stimulus set that is highly repetitive and potentially monotonous overall, which has drawbacks in terms of maintaining children’s attention. Experience has shown us that repeating short sentences with the same intonation contour on trial after trial causes children to lose interest sooner than when they have a more varied sequence of speech stimuli to listen to. One step in reducing this problem is to add a tag question or statement to the speech stimuli on each trial, following the sentence containing the target word sentence by a 1-s interval. These short sentences are all attention-getters such as *Can you find it*? or *Check that out!* The same speaker who produced the stimuli for the study records these tag sentences with lively intonation that varies from token to token, and they are appended to the speech stimuli in a way that maximizes variability from trial to trial by alternating sentences with rising and falling pitch contours. A second step in increasing variability is to intersperse test trials with filler trials, as described below.

### 2.4 Constructing stimulus orders

The visual and auditory stimuli are combined to construct two or more different stimulus orders. Although in earlier studies using this method we included only 10–20 trials in each experimental order, in more recent studies we have learned from experience how to get high quality data on a larger number of trials. One insight was the importance of incorporating the right amount of visual and auditory complexity in experimental and filler trials across the entire stimulus set, finding the appropriate balance that is optimal for a particular age group. By fine-tuning stimulus sets in this way, we can now test 18-month-olds on 30–40 trials, and older children on 40–50 trials. Trial types are counterbalanced so that the target object appears on the left and right sides an equal number of times, and does not appear on the same side more than three trials in a row.

In addition to using tag sentences to increase prosodic variability within experimental trials, it is also important to include filler trials that introduce visual as well as acoustic novelty across the stimulus set. If a study requires using multiple tokens of trials of one particular type, then the total number of trials may include up to 30% filler trials, designed to reduce repetitiveness and to maintain children’s interest. Filler trials might consist of jazzy pictures that are visually more complex and colorful than the experimental stimuli, such as a market scene or hot air balloons, accompanied by an exclamation such as *Good job - here’s another picture!* They might also consist of trial types we are pilot testing for other experiments, as long as the target objects are sufficiently different from those the children are being tested on. Thus one critical factor in maximizing children’s attentiveness throughout the experiment is to carefully balance variation and repetition across the stimulus set. Another critical factor is to provide a stimulus set that is challenging, but not too challenging, for children at each particular age.

### 2.5 Apparatus

The looking-while-listening procedure is conducted in a sound-treated room containing a test booth with two cloth-covered side walls, shown in Figure 1A. Built into the third wall of the booth is a polycarbonate rear-projection screen that is tough and scratch resistant. Behind the screen is a rear-screen projector designed with very short throw-space; thus the distance from the projector to the screen is only 1 m. Ideal for use in a small room. The video camera is positioned below the screen, mounted out of reach of the child. During testing, younger children sit in the booth on the parent’s lap on a swivel chair, with the seat height adjusted so that the child’s eyes are at the appropriate level in relation to the video.
puts the visual stimuli to the rear-screen projector in the testing room, and also to a monitor in the control room. The video record of the child’s eye movements on each trial is sent to a mixer in the testing room which integrates the video signal with graphic information about participant and trial number, stimulus order, and the onset and offset of auditory and visual stimuli. This combined information is then fed to a time code generator where it is time-stamped and sent to a VCR which records the complete video signal, including the child’s eye movements, the event information, and the time code. This composite video record is then digitized on a second computer in the control room.

2.6 Testing procedure

Each test session is preceded by a 15-min period during which the parent completes the procedures for informed consent and the experimenter interacts with the child. When both parent and child are comfortable and ready to begin, they are escorted to the testing room. The lights in the testing room are already dimmed, and images of puppets or cartoon characters familiar to young children are on the screen. A lamp mounted on top of the screen provides just enough light to avoid complete darkness, ensuring that the visual display is the most interesting thing in sight for the child. In a typical testing situation, two experimenters are involved: the first experimenter seats the parent and child in the testing room and checks that the camera is optimally positioned; the second experimenter, located in the control room, speaks to the child over the microphone to familiarize her with the sound source and directs attention to the visual display. If necessary, the second experimenter also coordinates with the experimenter in the testing room to focus the camera on the child’s face. Once the child is at ease and the parent has put on the darkened sunglasses, the first experimenter leaves the room and the test session begins.

On a typical test trial, the target and distractor pictures are shown for 2 s prior to the onset of the speech stimulus, providing the child with enough time to check out both pictures. As mentioned earlier, using each object as target in one trial and distractor in another trial constitutes an important control, offsetting the problem that one picture may be fixated first because it was named on a previous trial or simply because it is visually more salient. Figure 1B shows the time line for a typical trial. Trials last on average 6–8 s, including the 2 s before the onset of the speech stimuli, and are separated by an 800 ms brief interval when both screens are blank. The entire experiment lasts about 5–6 min.

In eyetracking studies with adults, participants are instructed to look at a central fixation point before responding to the verbal stimulus, and on most trials.
they willingly comply. This step has the advantage that the adult starts each trial at the same neutral location equidistant from all the visual stimuli, thus in principle a latency to shift from the center point to one of the stimulus objects can be assessed on every trial. A similar method has been used in eyetracking experiments with preschool children, who are requested to look at a central smiley face at the start of a new trial (see Trueswell this volume). In the looking-while-listening procedure we do not use a central fixation point, for the simple reason that infants and very young children will not follow such instructions. Although some preferential-looking studies have used a central fixation point at trial-onset, none have reported data on how long children actually maintain fixation at center. We explored this issue in a recent study by presenting a bright geometric image between the two stimulus pictures to 26- and 36-month-olds, in an effort to bring their attention to center at the start of the speech stimulus (Portillo, Mika, & Fernald 2006). However, within milliseconds children at both ages had shifted randomly to one side or the other before the target word was spoken. Since most children on most trials were already looking at the target or distracter object before hearing the target word, there is no evidence that a central fixation point is actually effective in studies with very young children.

2.7 Pre-screening for codeable trials

The video record of each test session is integrated with information about the participant (e.g., subject number, gender, age) as well as information relevant for data analysis (e.g., stimulus order and trial type) using the custom software Eyocoder. This information is not visible to observers during prescreening or coding. As shown in Figure 2, the video record of the test session also includes additional information that is visually accessible and useful to coders, such as the date, time code, subject number, trial number, and visual markers that appear at the points when the pictures and speech stimulus are presented to the child on each trial. Because coders are blind to trial type, they have no access to the visual stimuli seen by the child and they code eye movements in silence with no access to the verbal stimuli. Thus these markers are necessary for time-locking visual and auditory stimuli on each trial in relation to the child’s eye movements.

Pre-screening precedes coding and serves the primary purpose of flagging any non-useable trials on which the child’s response was affected by factors unrelated to the auditory and visual stimuli. Two pre-screeners, blind to the hypothesis as well as to side of target presentation, independently watch each testing session in real time, with the sound on in order to identify any talking by the child or parent during the trial. Because the pre-screeners have access to the verbal stimuli, they are never the same individuals who code eye movements for that study. The task of the pre-screener is to mark in the Eyocoder record any trials that should be eliminated from the analysis of the experimental trials for any of the following reasons: the child was not looking at the pictures prior to sound onset; the parent or child was talking during the trial; the child was fussy or inattentive; the child changed position so the face was not visible, etc. Only after both pre-screeners have agreed on all useable and non-useable trials does eye movement coding begin.

2.8 Coding eye movements

Coding involves reducing the continuous record of the child’s eye movements on each trial to a series of discrete events. Given that our goal is to measure the latency of the child’s shift in gaze from the distracter to the named target picture in relation to critical points in the speech stream, as well as the duration of fixations to the target and distracter pictures on each trial, high-resolution coding is essential. Highly trained coders, unaware of trial type and target location, use Eyocoder to move frame-by-frame through the digital record of each test session. The coder starts by synchronizing the time code with the onset of the speech stimulus, marking the frame to which subsequent events are referenced. Figure 3 illustrates the time line of eye movements captured in the Eyocoder record, from the onset of the utterance to the offset of the picture on two trials. The coder’s task is to
Table 1. Sample coding record for two 4-s trials using Eyecoder. Each line indicates the time at which the coder judged that a change occurred, either in the stimuli, e.g., from sound on to sound off, and/or in the position of the child's fixation, e.g., from right to off.

<table>
<thead>
<tr>
<th>Trial</th>
<th>Pictures</th>
<th>Sound</th>
<th>Fixation</th>
<th>Onset</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>on</td>
<td>on</td>
<td>right</td>
<td>20:32:22:02</td>
</tr>
<tr>
<td>2</td>
<td>on</td>
<td>on</td>
<td>off</td>
<td>20:32:53:05</td>
</tr>
<tr>
<td>2</td>
<td>on</td>
<td>on</td>
<td>left</td>
<td>20:32:23:10</td>
</tr>
<tr>
<td>2</td>
<td>on</td>
<td>on</td>
<td>off</td>
<td>20:32:24:17</td>
</tr>
<tr>
<td>2</td>
<td>on</td>
<td>on</td>
<td>right</td>
<td>20:32:24:22</td>
</tr>
<tr>
<td>2</td>
<td>on</td>
<td>off</td>
<td>right</td>
<td>20:32:26:20</td>
</tr>
<tr>
<td>2</td>
<td>off</td>
<td>off</td>
<td>right</td>
<td>20:32:26:21</td>
</tr>
<tr>
<td>3</td>
<td>on</td>
<td>on</td>
<td>left</td>
<td>20:32:30:10</td>
</tr>
<tr>
<td>3</td>
<td>on</td>
<td>on</td>
<td>off</td>
<td>20:32:32:09</td>
</tr>
<tr>
<td>3</td>
<td>on</td>
<td>on</td>
<td>right</td>
<td>20:32:32:13</td>
</tr>
<tr>
<td>3</td>
<td>on</td>
<td>on</td>
<td>off</td>
<td>20:32:32:26</td>
</tr>
<tr>
<td>3</td>
<td>on</td>
<td>on</td>
<td>left</td>
<td>20:32:33:04</td>
</tr>
<tr>
<td>3</td>
<td>on</td>
<td>off</td>
<td>left</td>
<td>20:32:33:28</td>
</tr>
<tr>
<td>3</td>
<td>on</td>
<td>off</td>
<td>off</td>
<td>20:32:34:02</td>
</tr>
<tr>
<td>3</td>
<td>on</td>
<td>off</td>
<td>right</td>
<td>20:32:34:56</td>
</tr>
<tr>
<td>3</td>
<td>on</td>
<td>off</td>
<td>off</td>
<td>20:32:34:23</td>
</tr>
<tr>
<td>3</td>
<td>on</td>
<td>off</td>
<td>left</td>
<td>20:32:34:28</td>
</tr>
<tr>
<td>3</td>
<td>off</td>
<td>off</td>
<td>left</td>
<td>20:32:34:29</td>
</tr>
</tbody>
</table>

3.2.1 Lab-wide versus study-specific reliability assessments

Lab-wide reliability checks ensure that no single coder deviates from the established coding norms of the entire lab. This involves selecting a single test session for all coders to judge independently, one in which the child shifts frequently or has significant head movement that makes eye gaze more difficult to judge. In other words, we select a test session in which the potential for disagreement among coders is high. Every coder is checked against every other coder to identify those individuals who may deviate from the lab standard by more than a single frame on any trial. If the reliability scores reveal a discrepancy, coders are re-trained to adhere to the norm, and any data they may have contributed since the last reliability check is reviewed and corrected as necessary. While lab-wide reliability tests examine agreement among all coders on a single testing session, study-specific reliability checks assess agreement among a subset of coders on several test sessions. This type of reliability assessment focuses on a specific data set to ensure consistency in coding within a particular study. Trials designated for study-specific reliability are chosen to provide the most rigorous reliability testing conditions for coders. We select 25% of the participants in a study for reliability checks, coding one quarter of those experimental trials on which two or more shifts occurred.

3.9.2 Entire-trial versus shift-specific reliability scores

The Eyecoder program enables comparison of the response line for each recorded event (i.e., picture and sound onset and offset as well as shifts in gaze), noting discrepancies that exceed one frame. This frame-by-frame comparison is summarized in two scores: an "entire-trial" percentage of agreement and a "shift-specific" percentage of agreement. The entire-trial agreement score is based on the percentage of frames on which two coders' judgments agree overall. Coders are considered to be in full agreement if they record an event on the same frame or differ from each other by no more than a single frame. This kind of overall agreement score is the traditional measure of reliability reported in many preferential-looking studies, and it may indeed be adequate when a global measure of preference is the major variable. However, an entire-trial agreement score does not provide a sufficiently rigorous measure of reliability when using more precise time course measures, given that the goal of such an analysis is to assess reliability at the same level of resolution as the measure of interest, in this case at the level of milliseconds. The problem is that the child is maintaining fixation on one or participate in regular reliability checks to assess the extent to which that all coders are "calibrated" according to lab standards. Coders are never informed that they are coding a session chosen for reliability purposes, to ensure that they are not being extra cautious because they know it is a reliability test.
the other of the two pictures on a majority of the 90 consecutive frames on each 3-s trial. Since the likelihood of agreement is much higher during these sustained fixations than during periods of transition, it is possible for two coders to receive a very high entire-trial agreement score, yet still disagree about the exact frames on which shifts in gaze begin and end. This is problematic because judgments about the initiation and duration of shifts are crucial in measures of reaction time and accuracy. Thus the traditional entire-trial score alone is too lenient a measure of reliability, although in combination with a more stringent reliability measure, it can help identify discrepancies among coders. Shift-specific reliability scores focus only on sequences of frames where shifts occur. The shift-specific calculation is a percentage of frames from shift start to finish on which coders agree, differing by one frame at most. If entire-trial and shift-specific scores are below 95% and 90%, respectively, an independent, highly-experienced coder examines the discrepancies between coders and makes a final decision about the gaze patterns for the session under review. At that time, coders who may have shifted standards for judging eye gaze are re-trained as necessary.

2.9.3 Data reduction using Datawiz
When coding is completed for all the participants in a particular study, the EyeCoder data from each coded test session needs to be coordinated with information about the identity and side of the target object on each trial, and the onsets and offsets of critical words in the speech stimulus. Using the custom software Datawiz, the EyeCoder data from all participants in the same study are consolidated, at the same time integrating other relevant information about each child such as vocabulary scores into the record. The output from the Datawiz program is an Excel-formatted spreadsheet for each child, with a series of codes indicating the time course of gaze patterns to the target picture, distracter picture, or away/off of both pictures at every 33-ms interval for every trial. These codes are aligned relative to pre-determined critical words in the stimulus sentence in each condition.

Depending on where the child is looking at the onset of the critical word, each trial is classified as target-initial, distracter-initial, off (between pictures), or away. Figure 4 shows a schematic representation of different types of response patterns over a series of hypothetical trials in which eye movements were time-locked to the onset of the target noun. The dashed vertical line indicates the average offset of a typical target noun, although in actual data this would vary from trial to trial, depending on the particular items used. The information shown in Figure 4 is comparable to the Datawiz output, except that the child's gaze patterns within each trial are summarized here using solid or patterned bars rather than as codes at every 33-ms frame. As indicated by the dark solid bars, trials 1 and 2 are both target-initial trials on which the child correctly maintained fixation on

![Figure 4](image_url)

Figure 4. Schematic representation of different types of response patterns on hypothetical trials in which eye movements are time-locked to the onset of the target noun. The dashed vertical line indicates average target-noun offset. This sequence of trials is meant to illustrate different classes of possible response patterns and is not representative of the actual distribution of trial types observed in test sessions with particular children.

the correct picture for at least 1400 ms after the onset of the target noun. Trial 9 is also a target-initial trial; however, on this trial the child shifted away incorrectly to the distracter picture after hearing about 433 ms of the noun. Trials 3 and 4 are distracter-initial trials on which the child shifted quickly to the target picture, yielding reaction times of 500 ms and 400 ms respectively. On trial 7, the child also started on the distracter, but this time the shift was initiated only 133 ms after noun onset. Such very short shift latencies are not included in the calculation of mean reaction time, because they are likely to have been initiated prior to the point where the child had enough acoustic information from the noun to make an informed response and then to initiate an eye movement. On trial 8, the child again started on the distracter picture, maintaining fixation on the wrong picture for the duration of the trial, without ever shifting to the referent. On other trials in Figure 4, the child was either between pictures or off-task at the onset of the noun. Please note that the sequence of trial types shown here does not represent
the distribution of response patterns for a typical experimental session with a child at an age for which the verbal and visual stimuli are appropriate. Trials 1–4 show the most common types of correct responses on distracter-initial and target-initial trials; trials 5–9 show patterns of correct and incorrect responding that are common, but occur less frequently, and trials 10–12 represent "away" trials that are relatively infrequent and would not be included in analyses of reaction time. Thus the goal in Figure 4 is not to show a representative sequence of trials for a particular child, but rather to provide examples of different gaze patterns that illustrate the variability possible in children's responses within the first 1800 ms following the onset of the noun.

2.9.4 Data cleaning
The output generated by Dataviz requires data "cleaning" steps before accuracy and RT can be calculated, to maximize the number of usable trials in the analyses. Recall that prior to the coding process, an experienced observer pre-screened each test session to exclude trials on which there was external interference or clear inattentiveness to the task. In the data cleaning process, the whole set of coded trials is further screened to evaluate off-task looking behavior within individual trials. The guidelines for accepting or rejecting individual trials are too detailed to describe here, but one example will suffice. On trial 3, the child started on the distracter picture and shifted to the target picture 500 ms from noun onset, in contrast to trial 11 where the child also started on the distracter and shifted away after 500 ms, but not to the target picture. In this case, the distracter-target shift in trial 3 would be included in the analysis of reaction time, while trial 11 would be excluded from the RT analysis because the shift in gaze was not directed toward either of the stimulus pictures.

3. Analyzing eye movement data from the looking-while-listening paradigm

This section examines different approaches to analyzing data on the time course of children's eye movements as they look at pictures and listen to speech that refers to one of the pictures. First we present plots of children’s shifting gaze patterns in relation to particular words in the stimulus sentence as it unfolds over time. Next we describe how to derive discrete measures of children's speech processing efficiency from the time course information for use in statistical analyses, focusing on reaction time and accuracy. And finally, we address the questions of how stable such time course measures of processing efficiency are within individual children, and whether they are meaningfully related to other dimensions of early language competence.

3.1 Plotting graphs of the time course of children's eye movements in relation to speech

A useful first step in examining data is to prepare an onset-contingent (OC) plot, which divides trials according to whether the child is looking at the onset of the critical word in the stimulus sentence. An OC-plot tracks separately the time course of participants' responses for target-initial trials and distracter-initial trials as the stimulus sentence unfolds. At the beginning of a trial, the child has no way of knowing which object will be named, and so is equally likely to be looking at the target or the distracter picture at the onset of the target word. Thus the behavior that constitutes a correct response varies with the position of the child's eyes at the onset of the target word: on distracter-initial trials, the child should quickly shift away from the distracter to the named target picture; however, on target-initial trials, the correct response is not to shift but to stay put. The OC-plot in Figure 5 provides a graphical overview of these two different response patterns, using data from one condition in a cross-sectional study with children in three age groups: 18-, 24- and 36-month-olds (Zangl & Fernald 2007). For each participant, trials were grouped contingent on which picture the child was fixating at the onset of the target noun. Plotted on the y-axis, at each 33-msec interval from target word onset, is the mean proportion of trials on which children at that point are looking at a picture that is different from where they started at target-word onset. Because OC-plots capture children’s tendency to shift away from the original starting point, they show both correct and incorrect responses: on distracter-initial trials, a shift away to the target picture is a correct response, while on target-initial trials, a shift away to the distracter picture is an incorrect response. Thus a child with perfect accuracy would shift quickly to the target picture on 100% of the distracter-initial trials, and would never shift away on target-initial trials.

The top three lines in Figure 5 track the mean proportion of distracter-initial trials on which children at each age have correctly shifted from the distracter and are now looking at the target at each 33-msec interval, plotted over participants from the onset of the target noun. The three lower lines track responses on target-initial trials for each age group, plotting the mean proportion of trials on which children have shifted away from the target at each time point and are now looking at the incorrect object. It is clear from the very different trajectories on distracter- and target-initial trials that the 36-month-olds were more likely than the younger children to respond correctly in both ways. On those trials when
they started out on the distracter picture, the 36-month-olds began shifting to the correct referent halfway through the target noun. However, when they happened to start out on the target picture, they tended to maintain fixation on the correct referent rather than shifting away. In contrast, the 18-month-olds responded less efficiently on both counts: on distracter-initial trials, their gaze patterns suggested they were more likely to shift more slowly and less reliably to the target picture, and on target-initial trials, they were much more likely to false-alarm by shifting away from the correct picture. If children were equally likely to make distracter-to-target shifts and target-to-distracter shifts in response to a particular object name, this response pattern would suggest they were unable to identify the target word and/or to match it with the correct referent. But to the extent that children are quick and reliable to shift to the correct referent on distracter-initial trials, and also tend to maintain fixation on target-initial trials without shifting away, we can infer that they are able to interpret the spoken word efficiently and have associated the name with the right object. By providing information on both types of correct response, the OC-plot in Figure 5 offers a global view of developmental

changes in children's efficiency in identifying the appropriate referent as the target noun is heard.

A simpler way to graph the same data is to use a profile plot, as shown in Figure 6. This plot tracks the mean proportion of looking to the target picture at each time interval, measured from target-noun onset, averaged over participants for the same three groups of children as in the previous analysis. The overall results in Figure 6 are, of course, the same as those shown in Figure 5, with the 36-month-olds responding more reliably and reaching a higher asymptote than the younger children. Note that the curves are relatively flat at the beginning of the target noun, before children had accumulated enough phonetic information to identify the word. But visual inspection of the slopes of the three curves, and of the relative points at which they begin to rise, suggests that the 36-month-olds as a group responded more quickly to the target word than did the younger children. In contrast to OC-plots, profile plots combine responses from both target- and distracter-initial trials to show the overall mean proportion of trials on which the child is looking to the correct referent at each time point, regardless of whether the child was already looking at the target picture or had just shifted there from the distracter.

Figures 7 and 8 provide another example of how OC-plots and profile plots provide different vantage points on the same data. The data represented in these figures come from a pilot study of children's use of semantic cues to identify the
referred to as an upcoming noun (Fernald 2004). Here we asked whether 26-month-olds, like adults (Altman & Kamide 1999), would respond more quickly to familiar nouns presented in sentence frames with a semantically related verb (e.g., *Eat the cookie*) than in frames with a semantically unrelated verb (e.g., *See the cookie*). Six object words were presented in both related and unrelated frames, carefully spliced to control for durations of carrier phrases and target words. Note that the graphs of children’s responses in these two conditions are plotted from sentence onset rather than from target noun onset, because we predicted that the very first word in the sentence would have an influence on their speed of orienting to the target object. Indeed, on related-frame trials, 26-month-olds began orienting to the referent much sooner than on unrelated-frame trials; the OC-plot in Figure 7 shows that responses on target- and distracter-initial trials began to diverge as the children heard the verb, before the target word had been spoken. In contrast, on unrelated-frame trials children had to wait for the target noun, since no earlier cues were available to facilitate identification of the referent. In Figure 8, the same data are presented in a profile plot, showing that children looked more overall to the target picture on related- than on unrelated-frame trials. Unlike the data on age-related changes in word recognition shown in Figure 6, there was no difference in asymptote between the two conditions here, because 26-month-olds eventually converged on the target object to the same extent on both kinds of trials. What the profile plot in Figure 8 shows instead is a substantial difference in the timing of children’s response between conditions. Like adults, 26-month-olds were able to take advantage of the semantic information in the verb to establish reference more quickly, but only on related-frame trials.

**3.2 Deriving measures of processing efficiency from time course data**

Although continuous plots of children’s eye movement data provide a dynamic picture of the time course of their responses to particular words in the unfolding sentence, they do not directly represent the discrete measures that are most convenient for purposes of statistical comparison. The two dependent measures used most frequently in our research are reaction time, or latency to orient to the target word, and accuracy, based on looking time to the correct referent calculated over particular regions of the sentence.
3.2.1 Reaction time

A ubiquitous measure in psycholinguistic research with adults, reaction time (RT) has been used to explore how different linguistic and non-linguistic factors influence speed of lexical access and ease of sentence interpretation. Many such studies with adults have used experimental paradigms, such as lexical decision, that not only require participants to make a voluntary behavioral response, but also depend on metalinguistic judgments. Given these task demands, RT measures have not been widely used in research with infants, who have a very limited repertoire of voluntary behaviors that can serve as reliable response measures. However, moving the eyes to interesting stimuli is one behavior with which infants have extensive experience, and developmental researchers have found many ways to use infant gaze as a revealing experimental measure (Haith 1980). Although in the 1980s, a few studies of infants' sensory and perceptual abilities used fine-grained temporal measures to investigate the early development of visual scanning (Aslin 1981; Bronson 1982), researchers interested in cognitive development used more global measures of infant gaze patterns to reveal "looking preference" to one stimulus over another. A decade later, researchers studying visual cognition began to use looking behavior to measure response latency. Investigating how infants develop expectations when looking at a display of alternating visual stimuli, Haith, Wempe, and Canfield (1993) established that the minimum latency for 3-month-olds to initiate a shift in fixation to a peripheral stimulus was around 200 ms. And when the infant had to disengage from one stimulus before initiating a shift to another, Hood and Atkinson (1993) found that responses were further delayed by 200 ms. These and other developmental studies of visual search (e.g., Canfield, Smith, Brezsnayk, & Snow 1997) were based on meticulous observations of infants' eye movements in a carefully controlled non-linguistic task, analyzed with millisecond-level precision. Coming from a completely different research tradition, these studies using visual RT measures with young infants converged with new research using eye movements to study spoken language understanding by adults (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy 1995), laying the foundation for the use of shift latency as a measure of processing speed in infant comprehension.

In the looking-while-listening procedure, RT is assessed on distracter-initial trials by calculating the latency of the infant's first shift away from the distracter toward the target picture, measured from a critical point in the stimulus sentence. Thus, the RT measure captures the point of departure from the initially-fixated distracter picture, not the point of arrival at the target picture, because we are interested in the decision to shift rather than the shift itself. Figure 9 shows the results of the first published study using RT measures of spoken word recognition by infants at 15, 18, and 24 months of age (Fernald et al. 1998). These results revealed that over the second year of life, during the same period when most infants show a "vocabulary spurt" in speech production, they also make dramatic gains in receptive language competence by increasing the speed with which they can identify familiar words and match them with the appropriate referent. These cross-sectional findings have now been replicated in a longitudinal study with English-learning infants (Fernald et al. 2006), and we have also found comparable results with infants from Latino families living in the US, learning Spanish as their first language (Hurtado, Marchman, & Fernald 2007).

In the Fernald et al. (1998) study, all distracter-to-target shifts that occurred between the onset of the target word and the end of the trial were included in the RT analysis. However, in subsequent experiments we have been more selective, excluding very short response latencies that presumably reflect eye movements programmed before the start of the target word. Every distracter-to-target shift that is interpretable as a correct response has multiple components, each requiring processing time that can only be estimated. These components include the time required to accumulate enough phonetic information to identify the spoken word and then to assess its relevance to the currently fixated picture, and the time required to initiate an eye movement, if a shift is required. Based on the findings of Haith et al. (1993), it seemed reasonable to assume that infants in the looking-
while-listening procedure need around 200 ms to program an eye movement, or perhaps longer given the need to disengage from an interesting distracter picture before shifting to the target (Hood & Atkinson 1993). Allowing time for 100 ms or so of phonetic information to accumulate, our first estimate was that it was reasonable to exclude from the RT analysis any shifts that occurred in the first 333 ms from the onset of the target word (Fernald et al. 2001). In more recent studies with older infants, we have used a cutoff of 300 ms (Zangl & Fernald 2007).

In addition to excluding very short latencies from the RT analysis, on the assumption that they were programmed before the target word was heard, it is also important to exclude very long latencies that are also unlikely to be in response to the target word. To develop guidelines for setting the boundary for such delayed shifts, we examined distributions of first shifts at different ages. The histograms in Figure 10 show the distributions of RTs on both distracter- and target-initial trials from a study of familiar word recognition by 18- and 21-month-olds (Fernald et al. 2001). Infants shifted to the target before the end of the trial on 88% of the distracter-initial trials; on 12% of these trials they never shifted at all. Note that the first shifts on six trials fell below the lower cutoff point of 333 ms, and thus were excluded from the RT analysis. In this data set we decided on 1800 ms as the upper boundary for shifts to be included in the RT analysis. This cutoff was chosen to eliminate responses influenced by the second repetition of the noun (since in this particular stimulus set each target word was spoken twice per trial), and it excluded outliers more than 2 SD greater than the mean of the distribution. For purposes of comparison, the distribution of response latencies on the 200 target-initial trials in this study is also shown in Figure 10B. Although infants shifted randomly to the distracter on some trials as the target word was spoken, more than half the infants did not shift at all, maintaining fixation on the correct picture.

The determination of appropriate cutoff points for excluding trials from an RT analysis is an important decision that may vary somewhat from study to study, depending on the experimental question and the age of the children in the study. The lower cutoff is relatively constrained; in eye-tracking studies with participants of different ages, this cutoff has varied from 200–400 ms (Bailey & Plunkett 2002; Ballem & Plunkett 2005), with shorter intervals typically used with adults (e.g., Tanenhaus, Magnuson, Dahan, & Chambers 2000) and with children older than 24 months (Fernald et al. 2006; Zangl & Fernald 2007). Establishing the upper cutoff is not always as straightforward, although one reasonable approach is to identify outliers by examining the distribution of shifts for outliers, as shown in Figure 10.

One additional issue of concern in calculating response latencies is the problem of sparse data, always a risk in infant studies that have very few trials. It is important to keep in mind that RTs can only be calculated on those trials where

![Figure 10](image-url). Distribution of RTs for first shifts from initially-fixated picture to alternative picture on 443 trials, following target word onset on (A) distracter-initial trials, and (B) target-initial trials. The proportion of trials on which no shift occurred within 3000 ms is noted on right (adapted from Fernald et al. 2001).

the child happens to start out on the distracter and then shifts to the target within the RT-window (e.g., 300–1800 ms from target-word onset), a subset of the data that often includes fewer than half the total number of trials. For this reason, in experimental designs with two or more within-subject conditions, it can easily happen that not all children contribute RT data and thus must be excluded from some analyses. For example, in a hypothetical experiment with 20 critical trials, 10 in each condition, each child will on average have eight usable trials, with two trials coded as “off” or “away” at target-word onset. Of the eight usable trials in each condition, some children may have six distracter-initial trials in each, all with shifts that occur within the appropriate RT-window. In this case the mean RT in each condition would be based on six trials — not an impressive number by the standards of studies with adults, but substantial for RT studies with infants. However, the more likely scenario is that children by chance will have only four...
distracter-initial trials in each condition, and some of these will be too fast or too slow to be included in the RT analysis. And of course some children by chance will have even fewer distracter-to-target shifts, and thus may end up with no RTs to contribute to the analysis at all. In earlier studies we have sometimes had to use mean RTs that were based on only two trials per condition, but this low criterion can result in very noisy data and works against finding positive results. To avoid this disappointing outcome, when designing an experiment in which latency measures are critical, it is important to make every effort to maximize the number of potential RT trials. This means including no more than two within-subject factors, since each additional factor reduces the number of distracter-initial trials in each condition that will potentially yield RTs. Another approach is to double the overall number of trials by observing each participant in two separate sessions, scheduled a day or so apart.

3.2.2 Accuracy

Accuracy reflects how reliably children look at the correct referent, operationalized as the mean time spent looking at the target picture as a proportion of total time spent on either the target or the distracter picture, averaged over a particular region of interest. While the RT analysis is based only on distracter-initial trials, accuracy includes both target and distracter-initial trials, assessing looking time to the referent regardless of whether the child started out on the target picture or had to shift to the target from the distracter. Depending on the experimental question, accuracy may be measured across a single broad time window, or over multiple smaller time windows. For example, when assessing developmental changes in infants' accuracy in recognizing familiar words in simple sentence frames (Where's the doggy?), accuracy was calculated as the mean proportion of looking to the target over the broad window extending from 300 to 1800 ms from the onset of the target word. This would be equivalent to the area under the curves over that window in the profile plot shown in Figure 6.

However, a multiple-window analysis was more appropriate for the study described earlier on children's use of semantic information from the verb to identify the referent (Fernald 2004). As shown in the profile plot in Figure 8, we defined four regions of interest: 1: verb, 2: determiner, 3: noun, 4: post-noun. Note that the measurement windows corresponding to these four regions incorporate the estimated 300 ms assumed to be necessary for processing the initial speech segments and mobilizing an eye movement; thus the verb window begins 300 ms into the verb and extends 300 ms beyond the offset of this word. Our prediction was that the accuracy curves on related-frame and unrelated-frame trials would begin to diverge at the end of the verb window, with significant differences emerging within the next few hundred milliseconds as children made use of informa-

![Figure 11](image.png)

**Figure 11.** Accuracy analysis: Mean proportion looking to target averaged over participants in four critical windows. Verb, Determiner, Noun, Post-noun. During Window 2 and Window 3, 26-month-olds fixated the correct picture significantly more on related-frame trials than on unrelated-frame trials (adapted from Fernald 2004).

...tion in the verb to find the target picture, even before they heard the noun. The accuracy analysis of these data is presented in Figure 11, showing the mean proportions of looking time to the target picture calculated over each of the four time windows. As predicted, the difference in accuracy between related and unrelated trials was significant in both the determiner and the noun windows, indicating that children could identify the correct referent using semantic information from the related verb. On related-frame trials, children were already fixating the appropriate referent 75% of the time, on average, by the beginning of the target noun; on unrelated-frame trials, in contrast, the mean proportion of looking time to the correct picture reached 75% only after the end the noun.

3.2.3 Stability and predictive validity of on-line processing measures

In the studies described here, on-line processing measures from the looking-while-listening procedure have been used to make between-group comparisons, tracking age-related differences in reaction time and accuracy in a cross-sectional design (Fernald et al. 1998), or examining condition differences in children of the same age (Fernald & Hurtado 2006; Thorpe & Fernald 2006) or between age groups (Lew-Williams & Fernald 2007). Using such group designs, we have shown that over the second and third year of life, children become faster and more reliable in recognizing familiar words in simple sentence frames, and that their ability to handle processing challenges such as morphosyntactic anomalies (Zangl &...
Fernald 2007) and more complex sentence structure (Thorpe & Fernald under review) also improves dramatically over this period. These investigations of on-line processing efficiency by very young language learners provide new insights into the early development of receptive language competence, complementing results from studies of lexical and grammatical growth that are based on more traditional measures of speech production over the first three years.

But characterizing typical patterns of language growth over time is just one perspective in developmental research; another central goal is to characterize variation among children. For example, young children vary widely in the size of their productive vocabulary, and one 15-month-old may produce more than 50 words while another has not yet started to speak at all, with considerable variability apparent in grammatical as well as lexical growth over the early years (Bates, Dale, & Thal 1994). Research using on-line processing measures of language understanding can address important questions about differences among children within an age group, as well as between-group differences. Is speed of processing at any given age a stable measure for an individual child? That is, are children who respond more quickly on average in identifying familiar words at 18 months, relative to the mean RT for children at that age, the same children who respond relatively more quickly at later ages? And how do individual differences in efficiency of spoken language processing relate to individual differences in language growth, as assessed by standard measures of lexical and grammatical knowledge? In particular, does processing efficiency in infancy predict language and cognitive outcomes at later ages?

To begin to address these questions, we conducted a longitudinal study of 59 English-learning infants, testing them in the looking-while-listening procedure at 15, 18, 21, and 25 months of age (Fernald et al. 2006). Children's speed and accuracy in spoken word recognition increased significantly over this period, consistent with earlier cross-sectional research. To explore the relation of on-line measures of speech processing skill to more traditional measures of linguistic development, parental reports of vocabulary and grammatical usage were gathered at five time points across the second year, along with a standardized test of lexical knowledge at 25 months. Speed and accuracy in speech processing at 25 months were robustly related to lexical and grammatical development across a range of measures from 12 to 25 months. Analyses of growth curves revealed that children who were relatively faster and more accurate in spoken word recognition at 25 months were also those who had experienced faster and more accelerated vocabulary growth across the second year.

These findings led to the obvious next question: to what extent do individual differences in processing efficiency in infancy predict later language and cognitive outcomes? In a recent follow-up study (Marchman & Fernald under review),

30 of the children from the Fernald et al. (2006) longitudinal study were tested at the age of 8 years on the Kaufman Assessment Battery for Children, Second Edition (KABC-II) and the Clinical Evaluation of Language Fundamentals, Fourth Edition (CELF-4) standardized assessments of cognitive and language skills. Multiple regression analyses were used to evaluate the long-term predictive validity of two measures in infancy — expressive vocabulary and mean RT at 25 months — in relation to school-age outcomes. In light of the presumed links between efficiency of spoken language comprehension and working memory in older children and adults, we also examined relations between processing speed in infancy and performance working memory. Mean RT in the looking-while-listening task at 25 months was significantly correlated with scores on CELF-4 (r = −.52 to −.55) and KABC-II (r = −.43 to −.70) at 8 years, taking vocabulary size into account, and relations were strongest to performance on the working memory subscale. Moreover, it was children's speed in identifying the target word on challenging trial types, requiring them to integrate semantic or morphosyntactic cues, that accounted for the most variance; mean RT to familiar words in simpler frames did not add predictive power. This prospective longitudinal study is the first to reveal the long-term predictive validity of on-line measures of processing efficiency by very young language learners, showing that individual differences in the efficiency of spoken language interpretation at the age of two years predict children's success in cognitive and language tasks in later childhood.

4. Conclusions

It is fascinating and revealing to watch infants look at the world as they listen to speech in a carefully controlled experimental context. Their gaze patterns provide a window on their referential decisions, as they seek meaning in the words they are hearing in relation to the objects they are looking at, all within fractions of a second. A great deal of research on early language development aims to characterize what words children "know" at a particular age, as if words were "acquired" in an all-or-none fashion; however, the methods and results we have described here take a different perspective, focusing on the gradual development of children's efficiency in using their emerging lexical knowledge to interpret spoken language. According to this view, if the rather static notion of "acquisition" is appropriate to lexical development at all, then learning to make sense of a spoken word is like acquiring a skill rather than acquiring a thing, with an emphasis on gradual mastery rather than on possession. Infants may respond to more and more words over the second year, but they also learn to respond with increasing speed and efficiency to
each of the words they are learning, and to recognize these words in more diverse and challenging contexts.

The looking-while-listening methodology described here uses fine-grained measures of the time course of children's gaze patterns in response to speech to explore the early development of language understanding. On the one hand, this procedure is technically "simple" in terms of stimulus presentation, similar to "preferential-looking" procedures in that it is low in task demands and does not require automated eyetracking technology. However, the looking-while-listening methodology differs critically from preferential-looking procedures in terms of the quantitative methods used for data reduction and analysis, yielding high-resolution real-time measures of speech processing rather than relying on summary measures of looking preference. Because gaze patterns are time-locked to the speech signal and coded frame-by-frame in this on-line paradigm, each 5-min experiment yields data from thousands of frames about the child's dynamic response to the unfolding sentence. As described in this chapter, the meticulous procedures involved in the collection, reduction, and multiple levels of analysis of such detailed data are certainly not simple. But they are well worth the effort, revealing a dynamic and nuanced picture of young children's developing skill in finding meaning in speech.

Acknowledgements

Thanks to Jonathan Berger, Nereyda Hurtado, Gerald McRoberts, Amy Purfors, John Pinto, Dan Swingley, and many other students and co-workers who have all contributed to the gradual refinement of the looking-while-listening paradigm described in this chapter. We are grateful to the National Institute of Child Health and Human Development (HD 42235) for their generous support of our research. Correspondence should be addressed to: Anne Fernald, Department of Psychology, Stanford University, Stanford, CA 94305.

References


