Available online at www.sciencedirect.com
ScienceDirect
Journal of Memory and Language xxx (2008) xxx–xxx
Journal of Memory and Language
www.elsevier.com/locate/jml

Analyzing 'visual world' eyetracking data using multilevel logistic regression ☆

Dale J. Barr

Department of Psychology, University of California, Riverside, CA 92521, USA

Received 14 November 2006; revision received 17 August 2007

Abstract

A new framework is offered that uses multilevel logistic regression (MLR) to analyze data from 'visual world' eyetracking experiments used in psycholinguistic research. The MLR framework overcomes some of the problems with conventional analyses, making it possible to incorporate time as a continuous variable and gaze location as a categorical dependent variable. The multilevel approach minimizes the need for data aggregation and thus provides a more statistically powerful approach. With MLR, the researcher builds a mathematical model of the overall response curve that separates the response into different temporal components. The researcher can test hypotheses by examining the impact of independent variables and their interactions on these components. A worked example using MLR is provided.

Keywords: Eyetracking; Statistics; Multilevel modeling

The current article provides solutions for the analysis of data sets from eyetracking experiments that use the 'visual world' paradigm (e.g., Cooper, 1974; Tanenhaus, Spivey, Eberhard, & Sedivy, 1995). In a typical 'visual world' experiment, participants view a scene and listen to speech containing references to objects in the scene. Such research is typically aimed at testing hypotheses of the form: "Does constraint X influence the processing of Y?" For instance, a researcher may wish to know whether verb-thematic information (Altmann & Kamide, 1999; Dahan & Tanenhaus, 2004) constrains referential search or whether the set of options available in a display constrains syntactic processing (Tanenhaus et al., 1995). Changes over time in the distribution of looks to elements in the scene are taken as an index of underlying linguistic processing. For example, consider the finding that listeners look more toward edible objects upon hearing the verb *eat* as in "the boy will eat..." than they do when the verb *eat* is replaced with the verb *move* (Altmann & Kamide, 1999). Such a finding suggests that verb information becomes quickly available and is used to derive expectations about upcoming referents.

The variable of time, inherently continuous, is an important variable in nearly every eyetracking experi-

☆ The author thanks Robert Rosenthal and the members of the weekly seminar on statistical methods (MAMA) at UC Riverside for their input on this project. One-on-one discussions with Dan Ozer and Chandra Reynolds were quite valuable in developing the framework. Thanks to Boaz Keysar, Kevin Esterling, Timothy Gann, Edmundo Kronmüller, and Mandana Seyfeddinipur for helpful comments; also to Peter Dixon, Len Katz, and two anonymous reviewers.

E-mail address: dale.barr@ucr.edu

0749-596X/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2007.09.002

ment. Spoken language is fundamentally a temporal phenomenon, and researchers are generally interested in examining how looking patterns change as evidence for various alternatives unfolds. The dependent variable (DV) in a visual world experiment is typically the region to which a participant directs his or her gaze at a given moment in time, a variable that is inherently categorical. Thus, the optimal analysis of such data sets calls for a framework that at the very least allows for the assessment of the effects of a continuous variable (time) on a categorical variable (gaze location).

The canonical statistical techniques used in experimental psychology—analysis of variance (ANOVA), the *t*-test, and related techniques—were developed for precisely the opposite situation: to assess effects of categorical variables (e.g., design factors) on continuous DVs (e.g., reaction time). Of course, analysts can transform their data sets to make them compatible with ANOVA. Time can be made categorical by breaking it up into a series of consecutive analysis windows and then performing separate analyses on each window (e.g., Hanna, Tanenhaus, & Trueswell, 2003; Kronmüller & Barr, 2007; Nadig & Sedivy, 2002). The categorical variable of fixation region can be made continuous by calculating a proportion that aggregates over time and over multiple trials (as is done in the vast majority of published studies).

However, there are costs to such an approach, some of which are often overlooked. Fortunately, a better solution is within reach. This article outlines these costs and offers a solution that uses multilevel logistic regression (MLR) with parametric curves. Although experimental psychologists often associate regression with observational rather than experimental data, all of the standard analyses performed using ANOVA—main effects, simple effects, and interactions—can performed in the MLR framework. Furthermore, the framework provides more flexibility for accommodating different kinds of predictor and dependent variables, be they continuous or categorical. It can straightforwardly handle continuous predictor variables in addition to time (e.g., participant covariates such as working memory span, or item covariates such as word frequency). The use of parametric curves gives the analyst more sophisticated ways of assessing change over time.

The components on which the MLR framework is built—multilevel modeling, logistic regression, and parametric curves—have all been addressed extensively in the statistical literature. Goldstein (2003), Raudenbush and Bryk (2002) and Snijders and Bosker (1999) among others give in-depth treatment to the topic of multilevel modeling. Readers who are unfamiliar with multilevel (also known as hierarchical or 'mixed effect') approaches may find it useful to consult any of various tutorial articles on the topic that have been written for an experimental psychology audience (Hoffman & Rovine, 2007; Quené & van den Bergh, 2004; Richter, 2006). Logistic regression is addressed in-depth in Agresti (2002), Cohen, Cohen, West, and Aiken (2003), and Hosmer and Lemeshow (2000), as well as in most textbooks on multilevel modeling. Finally, an accessible discussion of the use of parametric curves to model change over time can be found in Fitzmaurice, Laird, and Ware (2004).

Although information about the components of MLR is widely available, there is no one source that synthesizes information about these components in a format that is accessible to researchers working within an experimental framework. Many of the existing textbooks and technical articles are geared toward observational rather than experimental studies. Although the existing tutorial articles cited above on multilevel modeling present the topic in a manner that is accessible to experimental psychologists, they focus mainly on data sets where the response variable is continuous and for which modeling changes in the response over time is not of central interest. Furthermore, the two-level structure that is assumed in these articles is simpler than that required to deal with eyetracking data, where a minimum of three-levels will typically be required.

To give an overview, the article begins by discussing three problems arising from the application of conventional ANOVA techniques to visual world eyetracking data and shows how the various components of the MLR framework can overcome these problems. After laying out the MLR solution, it is then compared with other solutions on offer. Next, the framework is then demonstrated by means of a worked example, wherein the data from a published study are reanalyzed using MLR. Finally, the last 'Implementation' section of the article provides more in-depth technical details on the framework for researchers interested in applying it to their own data sets.

Note that although the article is geared toward the analysis of eyetracking data, the approach presented here could potentially be generalized to other applications involving repeated sampling of a response measure over brief intervals (e.g., motion tracking, monitoring movements of a computer mouse, etc.).

Accommodating time as a continuous variable using parametric curve regression

The main advantage of the visual world paradigm over more conventional psycholinguistic techniques, such as reaction time studies, is its potential to assess processing as speech unfolds over time. Unlike reaction time studies, in which the total information from a given experimental trial is concentrated into a single data point, eyetrackers sample participants' behavior repeatedly over very brief intervals, thus capturing the temporal dynamics of the underlying processing. Yet the potential to reveal underlying processing dynamics cannot be fully exploited within the conventional analytical approach, which requires time to be carved into discrete categories. In cases in which the researcher is only interested in gross patterns of change over time, such an approach may suffice. However, in many applications, discretizing time can lead the analyst to overlook important patterns in the data.

To illustrate, consider a hypothetical experiment examining pragmatic plausibility effects derived from verb semantics. In the experiment, listeners hear a sentence such as "the boy will swing the bat" while viewing a display containing several pictures. In one condition, one of the pictures in the display represents a pragmatically plausible object of the verb (e.g., a baseball bat). In another condition, this critical picture is replaced with a picture representing an object with a homophonous name, but that is pragmatically implausible as an object of the verb (e.g., a flying mammal bat). The hypothesis that we wish to test is whether pragmatic plausibility derived from verb semantics will constrain processing of the referential noun phrase "the bat". This hypothesis predicts a 'plausibility effect' on noun processing: during processing of the word "bat", looks should increase faster to a picture representing a pragmatically plausible object of the verb (e.g., the baseball bat) than to a picture representing an implausible object (e.g., the mammal bat).

Assume that the data from this experiment are subjected to a canonical analysis using ANOVA over proportions. The analyst time-locks the eye data to the onset of the noun "bat" and creates an analysis window spanning the entire word, which lasts 300 ms. Assume further that the window is shifted forward 200 ms to account for the approximately 200 ms that programming an eye movement requires (Matin, Shao, & Boff, 1993). So defined, the analysis window will maximally capture looking behavior driven by the on-line processing of the word "bat." The resulting proportions from various participants and items are then averaged together to yield the results presented in Fig. 1, panel (a). The ANOVA (or equivalently, *t*-test) finds a significant plausibility effect—that listeners were significantly more likely to look at the referent when it was a plausible object of the verb than when it was implausible (proportions of .6 versus .4). The researcher then concludes that listeners quickly derive information about pragmatic plausibility from verb semantics and use that information to constrain how they process the unfolding noun.

Such an analysis ignores the fact that the effect in (a) could have arisen from many possible underlying functions relating time to looking behavior. By collapsing

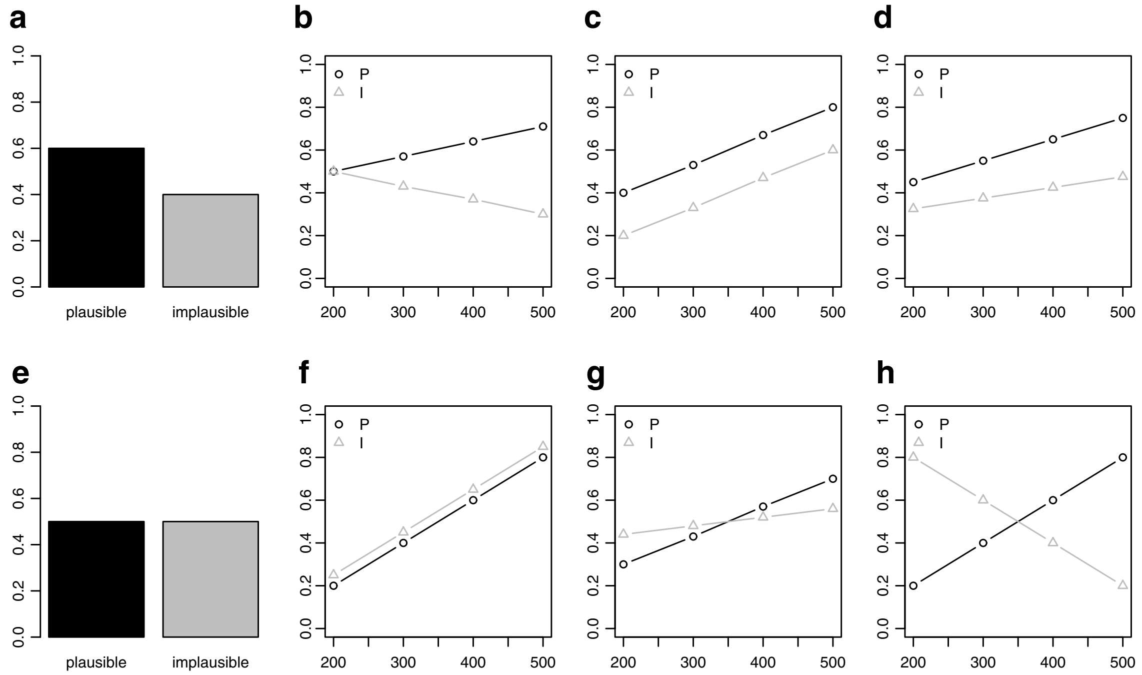

Fig. 1. Time-window analysis (a) showing a pragmatic plausibility effect and three hypothetical data patterns underlying the effect, (b)–(d). Panel (e) shows a null effect, and (f)–(h) show hypothetical patterns underlying it. The horizontal axis represents time from word onset, and the vertical axis represents the proportion of fixations.

time, the additional evidence that these patterns might provide in favor or against the researcher's hypothesis would be lost. Note that the logic of visual world experiments requires that the information instantiating constraint X be presented prior to the onset of the targeted linguistic fragment Y. Consequently, there are two different ways in which the constraint X can influence looking patterns. One way is by changing the distribution of looks *prior to the onset of the targeted linguistic fragment* Y, before perception of Y has even begun. Let us refer to such effects as an *anticipatory effects*. A second way that a constraint can have an effect is by modulating the way that fragment Y is processed. Such an effect would be evident in changes in the distribution of looks *as a function of time* during the period in which the fragment is being perceived and processed. Let us refer to these latter effects as *rate effects*.

Panels (b)–(d), for which the time variable has been uncollapsed, show three hypothetical ways in which anticipatory and rate effects can combine to yield the pattern in (a). Anticipatory effects are those which are present at the onset of the window; that is, any difference in height between the data points for each of the two lines at 200 ms. Note that it is logically impossible that any effect at the onset of the window would be the result of the processing of the noun, since processing of the noun has not started yet (again, assuming a latency of 200 ms for eye movement programming). Thus, the presence of an anticipatory effect can tell us that, indeed, the constraint had an effect on looking behavior, but it cannot tell us whether it has influenced how the noun was processed. Rate effects, in contrast, arise directly out of the processing of the targeted linguistic fragment to which the analysis window is time-locked. Such effects are given by the slopes of the lines over the window (keeping in mind that more complex, nonlinear functions of time are of course possible).

Claims about effects of some constraint X on the processing of Y can be made clearer and more compelling by distinguishing rate from anticipatory effects. The data in panel (b) shows a rate effect with no anticipatory effect: during processing of the noun, listeners became increasingly likely to fixate the plausible referent and decreasingly likely to fixate the implausible referent. Such a pattern would strongly support the interpretation that the processing of the noun was influenced by the semantic constraints of the verb. However, other patterns underlying (a) are possible that could weaken such an interpretation. For instance, panel (c) suggests an anticipatory effect of pragmatic plausibility, but no effect of plausibility on rate. To be sure, there is a main effect of time—as the word unfolds, listeners look increasingly toward referents that match the unfolding noun—but this increase is no greater for a picture representing a plausible object (baseball bat) than for one representing an implausible object (mammal bat). One might conclude from this that verb semantics caused listeners to anticipate pragmatically plausible referents, but did not modulate how the unfolding speech was mapped onto referents in the visual world. The pattern in (d) shows an anticipatory effect and a rate effect, but the rate effect is much smaller than that given in (b). Unlike in (b), where the likelihood of looking at the implausible object decreased over time, in (d) it increased, but at a slower rate than for the plausible object.

Now consider a different possible outcome of the experiment in (e), which suggests no difference between plausible and implausible objects. Such an outcome might suggest that the pragmatic plausibility constraint had no effect on noun processing. Such an interpretation, however, would be valid only if the underlying pattern was (f), where there is no anticipatory effect, nor any effect on rate—only a main effect of time. In (g), we see a rate effect, with a smaller slope for implausible than for plausible objects, that is altogether masked by an anticipatory effect in the opposite direction. Finally in (h) we see a similar masking of very strong rate effects by equally strong anticipatory effects in the opposite direction. The inverted anticipatory effect such as in (h) could have arisen out of a confound in the experimental materials; for example, because the pictures used for the implausible referents were more visually attractive than those representing the plausible referents.

In sum, the practice of collapsing time into discrete categories can obscure theoretically relevant data patterns by confounding anticipatory effects with rate effects. Without a clear separation of such effects, the analyst can be led to conclusions that are misleading or overlook rate information that is masked by an anticipatory effect. Certainly, researchers from different theoretical backgrounds may disagree over the theoretical significance of anticipatory versus rate effects. Nonetheless, whatever one's theoretical commitments, it is clear that effects that are present *prior to* the presentation of some targeted linguistic fragment should be given a different interpretation from those that arise *during* the processing of that fragment.

Deconfounding anticipatory and rate effects requires a framework that can more appropriately handle continuous variables such as time. To this end, we can appropriate sophisticated regression techniques for assessing change that have been developed for longitudinal studies. Even though the time span of an eyetracking study is many magnitudes smaller than in a typical longitudinal study, the interest in assessing change over time is identical.

One way of assessing change over time in longitudinal research is through the use of parametric curves (or 'growth curves') (Fitzmaurice et al., 2004). In the parametric curve approach, the analyst chooses an equation to represent changes in a dependent variable as a function of time. The general idea is to capture the overall pattern of change in a single function specified by several parameters. It is helpful to think of the function as modeling the grand mean response curve; that is, the mean response curve collapsed over all experimental conditions. One can then investigate how various experimental variables modulate the trajectory of the response via the parameters in the equation.

For example, the patterns observed in Fig. 1 can be captured using the equation for a line.

$$\eta = \pi_0 + \pi_1 t \tag{1}$$

The choice of the symbols η and π over the customary regression symbols y and β for the response variable and regression parameters (respectively) will be clarified in the Implementation section. The variable η represents the likelihood of gazing at a particular location (e.g., picture), and t represents time, as measured from the onset of the analysis window to which the eye data are time-locked. Typically, the data will be time-locked to the onset of a word or phrase, possibly adding a constant of 180–200 ms to account for eye-movement programming latency Matin et al., 1993. Note that the variables symbolized by π are the regression parameters (the fixed effects) that will be estimated in the analysis. Variable π_0 represents the intercept of the line, while π_1 represents the slope of the line.[1]

Note that when $t = 0$ (in other words, at the onset of the analysis window) η evaluates to π_0. Thus, the magnitude of π_0 tells us the likelihood of fixating the target region at the onset of the analysis window, thereby capturing anticipatory effects. In contrast, π_1 defines the rate effect, the rate of change in the likelihood of fixating the target region per unit time (e.g., per second). If π_1 is positive, that means that participants are becoming increasingly likely to fixate the target region; if π_1 is negative, that means that they were decreasingly likely to fixate the target region.

By using a multilevel analysis, discussed below, we can test for main effects or interaction effects of experimental variables on each of the regression parameters (the πs). In this way, we can distinguish between the effects of these variables on anticipatory versus rate components of the overall response. This is like performing two simultaneous ANOVAs, each providing a different picture of how the variables influence the underlying temporal dynamics of processing.

In some applications, the assumption of a strictly linear rate effect will be too simplistic, and a more complex function will be required. For instance, a linguistic signal may be temporarily ambiguous between a target (e.g., a bucket) and competitor object (e.g., a buckle), such that fixations to the competitor initially rise and then taper off once the point of disambiguation (e.g., the second syllable of “bucket”) is reached, resulting in an inverted “U” shape. The researcher can model the curvature by introducing a quadratic term into the equation, yielding

$$y = \pi_0 + \pi_1 t + \pi_2 t^2 \tag{2}$$

Ideally, the choice of the form of the function should be theoretically driven. When speech relatively unambiguously designates a target object, a simple linear function may suffice. When there is competition, inclusion of a curvature term may be necessary. Model selection techniques can additionally be employed to discover the simplest theoretically-motivated model that can fit the data. In the end, selection of an appropriate function will depend upon the goals of the analysis.

Accommodating gaze location as a categorical variable using logistic regression

Although linear regression is part of the general solution proposed here, there are problems directly applying such a framework when the dependent variable is categorical. In a visual world experiment, the dependent variable has as many categories as there are regions (or types of regions) in the presented scene. Such a variable will follow a multinomial distribution rather than a normal distribution (Agresti, 2002). The analyst is ultimately interested in understanding how the likelihoods of the participant’s gaze being in one of several states changes as a function of time and of the independent variables.

ANOVA, like linear regression, assumes that the dependent variable is continuous. It is customary to transform the categorical dependent variable into a continuous variable by calculating proportions. Proportions are computed by collapsing over time and over trials in the experiment. Although such a transformation makes the analysis tractable within ANOVA or linear regression, such an analysis violates the assumptions that the dependent variable has an unbounded range and that errors are distributed normally and independently of the mean. Proportions are bounded by the values 0 and 1, and error variance is proportional to the mean. Although the well-known arcsine transform of $2\arcsin\sqrt{y}$ (Howell, 1997) can stabilize the variance, the problem with analyses on proportional data sets goes beyond error variance or even potential floor or ceiling effects.

The crux of the problem is that effects on event likelihood are inherently *multiplicative*, while use of the proportional scale with ANOVA or linear regression assumes effects that are strictly *additive*. One way of seeing this is to note that ANOVA and linear regression assume constant effect sizes over the entire scale. How-

[1] Standard regression equations also contain a term representing residual error, but to simplify the exposition this term has been omitted.

ever, what is the meaning of an effect of .2 when the prior probability of the event is .8 or greater? Clearly, a probability greater than one is meaningless; thus, effects cannot be strictly additive, but must in some way depend on the prior likelihood of the event.

Statisticians often discuss event likelihoods in terms of odds. The odds of an event is defined as the ratio of positive occurrences (i.e., the event took place) to negative occurrences (i.e., the event did not take place). This contrasts with a probability, which is a ratio of positive cases to *all cases, positive or negative*.[2] The effect of some variable on the odds of an event are multiplicative; for example, a given variable might double or half the odds of the event. This assumption is evident, for instance, in epidemiological studies, where one typically finds statements of the form "exposure to substance S increased/decreased the odds of outcome Y by a factor of X." The critical phrase *by a factor of X* underscores the assumption of multiplicative effects.

To be able to apply conventional statistical techniques to likelihood data, one can take the log of the odds, which transforms multiplicative into additive effects. The log odds scale is the appropriate scale for assessing effects on a categorical dependent variable (see Agresti, 2002; Cohen et al., 2003; Jaeger, this issue, for further discussion). A proportion ϕ can be converted to log odds using Eq. (3).

$$\eta = \ln\left(\frac{\phi}{1-\phi}\right) \tag{3}$$

Analyzing data on a proportional scale instead of on a log odds scale can lead to improper estimation of effects. For example, Jaeger (this issue) shows how an analysis on a proportional scale can yield a spurious interaction effect when there are only main effects. With regard to eyetracking data, analysis on a proportional scale can also lead to misestimation of anticipatory and rate effects.

Consider the hypothetical data presented on the proportional scale in Fig. 2(a). Note that there is an anticipatory effect favoring one condition (triangles) over another (circles), and that the rate for the former condition also seems to be much higher. At 200 ms, the difference between the two lines on the graph is about .10; by 600 ms, the difference has increased to about .38.

However, this apparent rate effect turns out to be an illusion produced by the proportional scale. Consideration of the same data on the log odds scale in (b) shows that the two lines have parallel slopes, and differ only in anticipatory effect. The anticipatory effect amounts to 2 'logits' (units on the log odds scale). The slope for each line is .5 logits per 100 ms, which means that the odds of looking at the region increase by a factor of about 1.65 ($\exp(.5) = 1.65$) per 100 ms. The apparent rate effect disappears when the data are viewed on the log odds scale.

Assessment of the curves in Fig. 2(c) on the proportional scale suggests an anticipatory effect with no rate effect. However, plotting the same curves on the log odds scale indicates a steeper initial increase in one condition (triangles) than in the other (circles).

The appropriate technique for analyzing these data is either logistic regression, or linear regression with an appropriate logistic transformation of the response (Agresti, 2002; Jaeger, this issue; McCullagh & Nelder, 1989). As it happens, whenever the proportions lie between .3 and .7, logistic regression will tend to yield the same results as will linear regression on a proportional scale. Yet for many eyetracking data sets, proportions will fall below this range (e.g., due to fixation on a central fixation cross or the fact that gazes are distributed over a large number of regions in a display). In the long run, an approach based on log odds will be more widely applicable.

Note that parameter estimates that are obtained in logistic regression are on the log odds or logit scale, and thus represent the log odds of the target event (e.g., looking at a particular object). The logit scale is unbounded and symmetric around zero. A logit of zero means that the target event was equally likely to occur as not to occur (e.g., a probability p of .5). When the logit is positive, the target event is more likely to occur than not ($p > .5$); when it is negative, the target event is less likely to occur than not ($p < .5$). For instance, with a logit of 1, the odds of the target event occurring are 2.71 times ($\exp(1) = 2.71$) the odds of not occurring. A logit of -1 indicates that the target event is 2.71 times more likely *not to occur* than to occur.

In the rest of this article, it is assumed for the sake of simplicity that the response variable has only two categories (i.e., that it is "dichotomous"). In reality, the response variable for most eyetracking experiments will have three or more categories (i.e., will be "polytomous"). For instance, a given experiment might contain displays with three objects: a target, a competitor object that is related to the target in some way, and an object that is unrelated to the target. A generalization of logit models, known as multinomial logit models, can be used to analyze all categories of a polytomous variable. With existing software, however, it is difficult to fit such models. Standard logistic regression can be used if the polytomous variable is first converted into a dichotomous variable, by collapsing all but one of the categories into a single group (e.g., target versus "other").

Accommodating nonindependence of observations using a multilevel approach

Up to this point, I have indicated the need for a logistic regression framework; the 'logistic' part is required to

[2] Given a probability p, one can compute the odds as $\frac{p}{1-p}$; given an odds q, one can compute the probability as $\frac{q}{1+q}$.

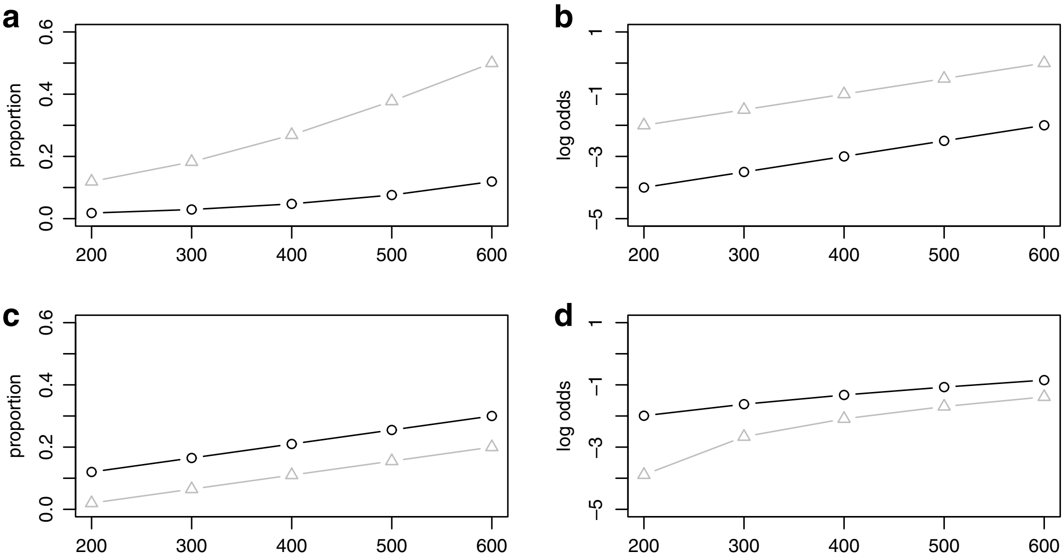

Fig. 2. Hypothetical response curves on the probability scale (left panels) and on the log odds scale (right panels).

accommodate a categorical dependent variable, and the 'regression' part is needed to accommodate the continuous predictor variable of time. But there is an additional requirement that the framework must also meet: namely, it must be able to handle the fact that not all of the observations that comprise an eyetracking data set are independent. The nonindependence of observations in visual world experiments derives from two sources: (1) the fact that such experiments have a multilevel sampling scheme; and (2) mechanics of how the eye moves. It is important to take these sources of clustering into account when statistically analyzing eyetracking data sets. Treating observations as independent when they are not can lead to underestimation of standard errors, which in turn can inflate the Type I error rate.

Eyetracking experiments have a multilevel sampling scheme in that a single trial of an experiment yields not one, but many observations. Eyetrackers sample the position of the participant's eye gaze repeatedly at fixed time intervals (in most cases, about once every 17 ms). The observations collected during a given trial will be more correlated with one another than observations collected during different trials. Unlike a typical reaction time experiment, one can have tens, hundreds, or even thousands of observations for a given presentation of a particular experimental item to a given participant. Modeling time using parametric curves can make the observations within trials conditionally independent from one another given the model (Cohen et al., 2003; Fitzmaurice et al., 2004), but there are other sources of nonindependence that must also be addressed.

Except in the very unlikely case of an experiment that includes repeated observations on only one participant, most designs will include higher order sources of clustering (non-independence); namely, the nesting of trials within participants and/or experimental items. The sampling hierarchy for a standard experiment can be captured in a three-level model (Fig. 3): level-1 is the level of individual observations (individual data frames), level-2 is the trial level (trials in the experiment), and the level-3 units are both participants and items (each level-2 unit is "cross-classified" at level-3 by participant and item; see Raudenbush & Bryk, 2002). Each participant forms a unique 'cluster' in which observations are correlated, and each item also forms such a unique 'cluster.' If we assume that each item is presented to each participant only once—as is the case in most psycholinguistics experiments—then each trial in an experiment is a unique combination of the two clusters.

The standard solution to this problem in experimental psychology is to aggregate data up to the highest level in the sampling hierarchy. For example, mean proportions are computed for each participant (or item) in an experiment, and these independent means, instead of the original observations, are submitted to the analysis. Such a solution can achieve independence of observations, but at the cost of losing information and thus, statistical power.

In contrast, multilevel (or 'mixed-effect') regression can solve the problem of nonindependence while incurring no such cost. Multilevel models solve the problem by directly modeling nonindependence through the inclusion of 'random effects' corresponding to the various clusters in the sampling design (Raudenbush & Bryk, 2002). This largely avoids the loss of information due to data aggregation, and thus will generally improve statistical power.

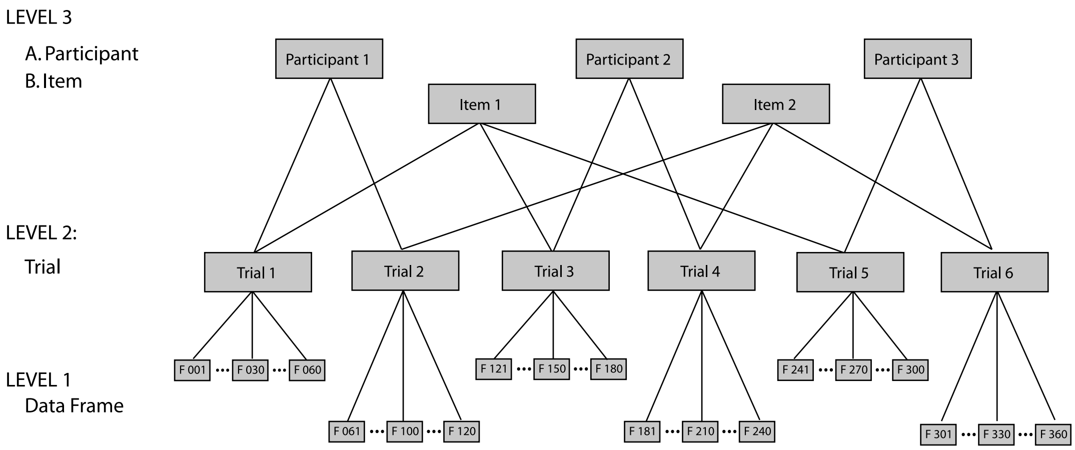

Fig. 3. The sampling hierarchy in a typical eyetracking experiment.

The second source of nonindependence in eyetracking data comes from the way that the eyes move. Consecutive frames in an eyetracking data set may be more or less dependent, depending on what they eyes are doing at that time. It is not physically possible for a participant's eye gaze to instantaneously travel from one region to another; the gaze must travel through time and space to reach its destination. Yet true independence presupposes that such instantaneous shifts are possible. Furthermore, it might be the case that the gaze is unlikely to shift while the planning of an eye movement is underway, such that observations that take place during eye movement planning will be correlated. Ideally, one would want to somehow incorporate these potential sources of nonindependence into the model, but it is not immediately clear how this could be accomplished.

Failure to take into account eye-movement based dependencies can lead to underestimation of the standard errors that are used in hypothesis testing. Two solutions are recommended for overcoming these dependencies. The first is to compute 'robust' or 'Huber-corrected' standard errors instead of the regular standard errors. These robust standard errors provide protection against Type I error inflation due to model misspecification (Raudenbush & Bryk, 2002). However, they are generally only appropriate for experiments involving many participants and items.

Alternatively, one can filter out these dependencies by aggregating together all trials within a given condition, and grouping observations into a series of temporal bins (e.g., 50 ms bins). One can then compute the 'empirical logit' for each bin, which is a quasi-logit transformation that is designed to handle cases for which the standard logit is exceedingly large or small (e.g., when the probability is near zero or near one).[3] Through such aggregation, one can filter out the eye-movement based dependencies, but at the cost of a loss of statistical power. Nonetheless, such an approach will be appealing when one is dealing with data sets including a relatively small number of participants or items.

Comparison to existing approaches

The MLR solution advocated here is intended as a general solution for analyzing data from eyetracking experiments. The regression framework enables treatment of time as a continuous variable, while the logistic transformation accommodates the categorical DV of gaze location. Furthermore, the problem of nonindependence due to repeated sampling is handled through multilevel modeling, and eye-movement based nonindependence is resolved through either robust standard errors or aggregation using the empirical logit.

Other solutions that have been proposed bear some similarities to the MLR framework, although only the MLR framework simultaneously handles the continuous variable of time, the categorical variable of gaze location, and eye-movement based dependencies in the data. A similar solution was proposed by Magnuson, Dixon, Tanenhaus, and Aslin (2007), who also used a growth curve (i.e., parametric curve) approach. But in contrast to MLR, they analyzed their data on a proportional scale. Furthermore, they fit curves using orthogonal polynomials, rather than the natural polynomials used here. With orthogonal polynomials, the interpretation of each term in the equation is independent of all other terms (i.e., inclusion of a higher-order term does not change its interpretation). Thus, the intercept term gives the mean height of the curve over the entire analysis window, unlike in the MLR approach, in which it corresponds to the mean difference at the intercept. The orthogonal polynomial approach can in fact be used

[3] See the Implementation section for details on computing the empirical logit.

within MLR, and would be most appropriate in situations in which there is little need to control for anticipatory effects. However, when such effects are present, then the orthogonal polynomial approach will tend to confound them with rate effects.

Knoeferle, Crocker, Scheepers, and Pickering (2005) handled the categorical nature of eye movement data using a hierarchical loglinear regression model. However, loglinear models, unlike logistic regression models, can accommodate only categorical predictor variables, and thus cannot incorporate time as a continuous variable. Similarly, Arai, van Gompel, and Scheepers (2007) used a log-ratio technique in order to compare fixations between two regions of interest. The DV was the log of the ratio of proportions between the target region and a competing region. But as in this former study, the time variable was collapsed into a series of discrete time windows, each represented by a mean log ratio. The log ratio and log linear approaches may be most useful when the goal of the analysis is to compare across regions, and when the need to separate anticipatory and rate effects is not critical.

Scheepers, Keller, and Lapata (in press) fit an 11-parameter Logistic Power Peak function to their data. Although the function had a logistic form, the data that were fit consisted of differences calculated on the proportional scale. Furthermore, due to the complexity of the fitting process, the function was fit individually to each of eight groups (rather than being fit to individual participants or items), and thus the standard errors of the estimates did not correspond to any traditional sampling unit. In contrast, MLR provides the customary standard errors computed over participants and items. Furthermore, unlike in a sequential curve-fitting approach, the MLR approach estimates fixed and random effects simultaneously, thus minimizing biases in the estimation procedure (Raudenbush & Bryk, 2002).

Worked example: re-analysis of Kronmüller and Barr (2007)

In this section, the MLR approach is illustrated by way of an example. A subset of the data from Experiment 2 of Kronmüller and Barr (2007) will be reanalyzed using the MLR approach. More in-depth details regarding the implementation of the approach can be found in the final Implementation section of the article.

In the experiment, participants viewed a screen displaying pictures of three unusual objects. In each trial, participants listened to a person (the 'speaker') describe one of the three objects (the *target* object) and attempted to identify the target based on the description. For example, the listener would hear a description such as *the flying saucer* and then would select the best-matching object by pressing a button. Listeners' eyes were tracked as they searched for the target object.

The data considered here were drawn from a subset of the design in which the expression used to refer to the target had been used previously to refer to the same object (i.e., the *Maintain Precedent* condition in the original article). The previous mention of the object was either by the same speaker as was currently speaking (*Same Speaker* condition) or by a different speaker (*Different Speaker* condition). Additionally, listeners were either under cognitive load or not under load when they interpreted the expression.

Kronmüller and Barr (2007) found that comprehension was facilitated when an expression was re-used, but found no evidence that this facilitation was any greater in the Same Speaker condition than in the Different Speaker condition. This is consistent with earlier studies eyetracking studies that also report speaker-independent facilitation for repeated descriptions (Barr & Keysar, 2002; Metzing & Brennan, 2003). However, it is inconsistent with another literature concerning memories for spoken words which finds speaker-specific facilitation due to perceptual priming (e.g., Church & Schacter, 1994; Goldinger, 1996; Nygaard & Pisoni, 1998). This literature finds that people are faster to identify or recognize a word when it is spoken by the same speaker than when it is spoken by a different speaker. It is surprising not to find this priming effect in the context of a referential communication experiment.

One reason why such a perceptual priming effect may not have been detected is because the analysis that was used was insufficiently sensitive. Kronmüller and Barr (2007) used a sequential time-window analysis, dividing the data into a series of 300 ms bins, and performing an ANOVA on each bin. It is possible that the priming effect may have been too short-lived, or may have been broken up across subsequent windows. Therefore, we return to these data in order to determine whether the enhanced sensitivity of MLR leads to the detection of a speaker-specific priming effect.

The data set for Experiment 2 of Kronmüller and Barr (2007) consisted of data from 56 participants over 32 experimental items. Half of these items were not in the Maintain condition, and were therefore excluded from the following analysis. Two trials were excluded due to problems with the experimental procedure. Eye data was recorded at a rate of 60 frames per second (one sample approximately every 17 ms). For simplicity, we consider data for a time-window of approximately 300 ms (the data shown in Fig. 5).

We begin with a conventional analysis of the data such as was conducted by Kronmüller and Barr (2007). The data in Fig. 4 shows the proportions of fixations to the target during an analysis window spanning from 180 ms after the onset of the referring expression to 450 ms. A conventional 2 × 2 repeated-measures

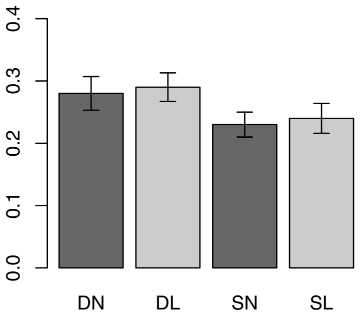

Fig. 4. Results from Kronmüller and Barr (2007), collapsed over a 180–450 ms time window. DN = Different Speaker, No Load, DL = Different Speaker, Load, SN = Same Speaker, No Load, SL = Same Speaker, Load.

ANOVA was conducted on the arcsine-transformed proportional data. This analysis revealed a main effect of Speaker, in which there was a higher likelihood of fixating the previously-mentioned target when the speaker was different (.28) compared to when the speaker was the same (.23), $F1(1,55) = 3.94$, $MSE = .004$, $p = .05$; $F2(1,31) = 6.13$, $MSE = .090$, $p < .05$. Note that this effect is in the opposite direction of what one would expect under the hypothesis of speaker-specific priming: listeners were slightly *more* rather than less likely to look at the target when it had been previously mentioned by a different speaker than the one currently speaking.

Instead of attempting to interpret what such an effect might mean, let us consider the uncollapsed data (Fig. 5).[4] It is clear from this figure that the difference detected by the previous analysis is entirely driven by a strong anticipatory effect at the onset of the window (180 ms), since the rate effect appears to go in the opposite direction.

How could this anticipatory effect be explained? Note that in the experiment, speakers' utterances were blocked, so that participants knew who the speaker would be before any given trial began. Thus, it was possible for them to think about what objects the speaker had mentioned before and to try to anticipate what the speaker would refer to next. In the Same Speaker condition, the speaker had previously mentioned two of the three objects in the display; in the Different Speaker condition, two of the objects had been mentioned as well, but by the old speaker; the current speaker had mentioned none of them. The proportion of looks to regions in the display other than the target provides evidence that the lower rate of looks to the target in the Same Speaker condition was driven largely by more looks to the one (non-target) object that the speaker had not yet mentioned (.37). This suggests that they (wrongly) anticipated that this previously unmentioned object would be the next target, hence the lower rate of looks to the actual target. In contrast, listeners in the Different Speaker condition gazed at that object no more than at the other two objects (.31), thus distributing their gaze roughly equally over the objects in the display.

It is critical to note that Fig. 5 not only shows this strong anticipatory effect favoring the target in the Different Speaker condition, but also suggests a rate effect that is in the predicted direction of the speaker-specificity hypothesis, with a steeper increase in the Same Speaker condition than in the Different Speaker condition. Although the curves in the Same Speaker condition start off lower than those in the Different Speaker condition, by the end of the window, they are at par. Furthermore, the rate effect does not appear to be modulated by cognitive load, thus corroborating the hypothesis of an underlying priming mechanism. This rate information was completely lost in the conventional analysis because of the need to collapse over time.

Re-analysis using MLR

The model that was chosen for the analysis was a simple linear function, with the onset of the analysis window time-locked at 300 ms after the onset of the critical word. The 300 ms point was determined by plotting the grand mean and visually assessing the earliest rise in fixations to the target (Fig. 6). Using the grand mean is a conservative approach to determining time-locking in that it is blind to condition. Note that the earliest point at which one could have possibly observed gaze behavior driven by processing of the linguistic signal would be at 180–200 ms after word onset, due to the latency for eye movement planning (Matin et al., 1993). The fact that signal-driven gaze behavior does not appear until about 100 ms later conforms to the fact that the descriptions of the objects in the experiment were unfamiliar and referred to using unconventional descriptions. Such expressions are intrinsically more difficult to interpret than references to familiar, everyday objects using conventional names.

The model that was chosen was a simple linear model, given by the following equation:

$$\eta = \pi_0 + \pi_1 t. \qquad (4)$$

In this equation, η represents the estimated log odds of fixating the target for a given frame and t represents the time elapsed from the onset of the analysis window. The variables π_0 and π_1 will be estimated by the regression analysis, and will capture any anticipatory and rate effects, respectively.

[4] The fact that the curves look similar on the log odds and proportional scale is because they occupy the range of the scale that is roughly linear, namely, between about .3 and .7 (Agresti, 2002; Jaeger, this issue).

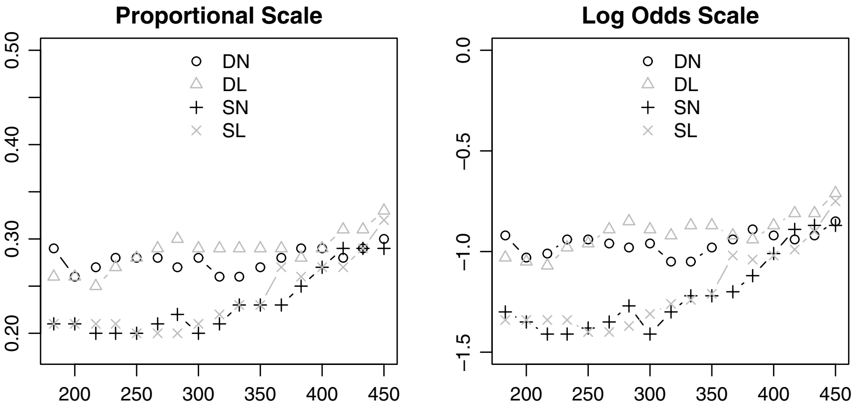

Fig. 5. Results from Kronmüller and Barr (2007) shown on proportional and log odds scales. DN = Different Speaker, No Load, DL = Different Speaker, Load, SN = Same Speaker, No Load, SL = Same Speaker, Load.

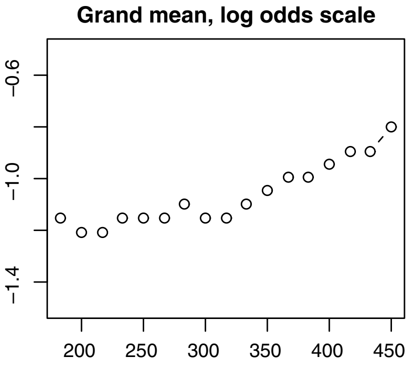

Fig. 6. Mean fixations to the target object collapsed over conditions, log odds scale.

Using the multilevel framework, it is possible to examine how the experimental variables of Speaker (Same versus Different) and Cognitive Load (Load versus No Load) and their interaction may have generated an anticipatory effect or modulated a rate effect. The essence of the analysis is that each of the two level-1 π parameters are modeled, in turn, at levels 2 and 3 in terms of the fixed effects of the experimental variables and the random effects of trial and participant/item.

Estimation of regression coefficients

Two separate multilevel analyses were performed, each demonstrating a different way of controlling for eye-movement based dependencies in the data. The first analysis was multilevel logistic regression (MLR) with robust standard errors. This analysis was performed on the full data matrix, with no aggregation. The second analysis was weighted empirical logit regression (quasi-MLR), which controls for such dependencies through aggregation. The data were aggregated into a sequence of four 50 ms bins for each cell of the design for each participant. For both the logistic regression and empirical logit analyses, separate participant and item analyses were performed.[5]

The two analyses yielded estimates of the intercept (π_0) and slope (π_1) terms for each condition (Table 1). Observed values are plotted against model fits in Fig. 7. It is evident that the models capture the overall trends in the data quite well.

For the logistic regression analysis, t values output by the HLM software are reported for both the participant (t_1) and item (t_2) analyses. The analysis detected a significant main effect of Speaker on the intercept (π_0) [$t_1(890) = 2.73$, $p < .01$; $t_2(890) = 2.68$, $p < .01$]. Specifically, the log odds of fixating the target were .32 higher in the Different Speaker than in the Same Speaker condition, with condition means of −.93 and −1.25 respectively. In other words, the odds were higher in the Different Speaker condition by a factor of about 1.37 (exp(.32) = 1.37), indicating an anticipatory effect of the Speaker manipulation. There was no main effect of Load [$t_1(890) = .43$, $p > .5$; $t_2(890) = .42$, $p > .5$] nor Speaker by Load interaction [$t_1(890) = .13$, $p > .5$; $t_2(890) = .09$, $p > .5$].

For the slope term, there was a clear main effect of time, with looks to the target increasing at a rate of 1.93 logits per second [$t_1(55) = 5.12$, $p < .01$; $t_2(31) = 5.56$, $p < .01$]. In other words, for every 100 ms, the odds of fixating the target increased by a factor of 1.21 (exp(.193) = 1.21). The critical question is whether this rate effect was modulated by the Speaker variable, with a higher rate in the Same Speaker condi-

[5] See the General Discussion and Implementation sections for discussion of why separate participant and item analyses were required.

Table 1
Parameter estimates (based on the analysis by participants)

Parameter	Speaker	No Load		Load		Collapsed	
		Est.	*SE*	Est.	*SE*	Est.	*SE*
Logistic regression							
Intercept (π_0)	Different	−.90	.12	−.97	.14	$-.93^{\mathbf{A}}$	.10
	Same	−1.23	.13	−1.27	.09	$-1.25^{\mathbf{A}}$	.07
	Collapsed	−1.06	.09	−1.12	.09	−1.09	.06
Slope (π_1)	Different	.90	.93	.85	.78	$.87^{\mathbf{B}}$	.53
	Same	2.89	.88	3.09	.66	$2.99^{\mathbf{B}}$	.50
	Collapsed	1.86	.66	1.95	.58	1.91	.38
Empirical logit							
Intercept (π_0)	Different	−1.16	.18	−1.28	.18	-1.22^{c}	.12
	Same	−1.58	.18	−1.50	.18	-1.54^{c}	.12
	Collapsed	−1.37	.13	−1.39	.13	−1.38	.09
Slope (π_1)	Different	1.34	1.11	1.70	1.11	$1.52^{\mathbf{D}}$	.78
	Same	4.22	1.11	3.26	1.11	$3.74^{\mathbf{D}}$	.78
	Collapsed	2.78	.79	2.48	.30	2.63	.56

Note: Bold capital letters indicate significance ($p < .05$).
Lowercase letters indicate marginal significance ($p < .10$).

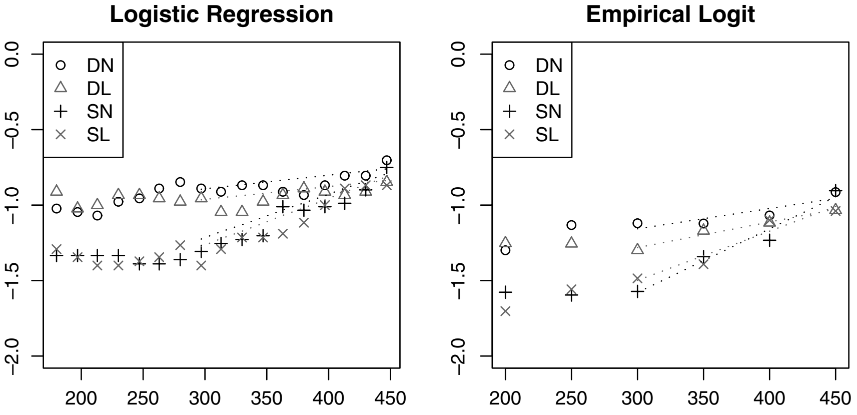

Fig. 7. Observed values (symbols) and model fits (dotted lines). The data are presented on the log odds scale.

tion than in the Different Speaker condition. This hypothesis was supported: the estimated slope in the Same Speaker condition was 2.99 logits per second, 2.12 logits higher the estimate of .87 logits per second for the Different Speaker condition [$t_1(890) = 2.97$, $p < .01$; $t_2(890) = 2.43$, $p < .05$]. This implies that for every 100 ms, the odds of fixating the target increased by a factor of 1.34 in the Same Speaker condition, compared to a factor of 1.09 in the Different Speaker condition. There was no main effect of Load on the slope [$t_1(890) = .08$, $p > .5$; $t_2(890) = .11$, $p > .5$]. More importantly, the Speaker effect was not modulated by the Load manipulation [$t_1(890) = .16$ $p > .5$; $t_2(890) = .16$; $p > .5$], supporting the hypothesis that the Speaker effect was due to perceptual priming.

The multilevel empirical logit regression was conducted using the software package R (R Development Core Team, 2007), with package lme4 (Bates, 2007) to obtain parameter estimates and package languageR (Baayen, 2007) to obtain *p*-values. The 'lmer' function yields *t* statistics but the degrees of freedom are not estimated. To obtain *p*-values, a Markov Chain Monte Carlo (MCMC) method was used (Baayen, Davidson, & Bates, this issue).

The empirical logit analysis revealed basically the same pattern of results, although the analysis was less powerful. For the intercept, the main effect of Speaker was marginally significant [$t_1 = 1.82$, $p = .067$; $t_2 = 1.83$, $p = .067$]. There was no main effect of Load [$t_1 = .13$, $p > .5$, $t_2 = .19$] nor any interaction effect

[$t_1 = .59$, $p > .5$, $t_2 = .40$]. For the slope, the main effect of Speaker was significant [$t_1 = 2.00$, $p < .05$; $t_2 = 2.13$, $p < .05$] and in the predicted direction, while the main effect of Load was not significant [$t_1 = .27$, $p > .5$; $t_2 = .20$, $p > .5$]. The interaction was also not significant [$t_1 = .59$, $p > .5$; $t_2 = .26$, $p > .5$].

In conclusion, the MLR framework can discover theoretically important patterns in eyetracking data that can go undetected using conventional analyses that collapse over time. A conventional analysis of the data using ANOVA found a main effect of Speaker, with listeners more likely to look at the target in the Different Speaker condition than in the Same Speaker condition. However, an analysis using MLR revealed this difference to be entirely driven by an anticipatory effect; that is, listeners were less likely to look at the target prior to the onset of the referring expression in the Same Speaker condition, apparently due to a slight anticipation that the speaker would refer to a new referent and not the target. As the speech unfolded, listeners in both conditions looked toward the target, but the rate of increase was stronger in the Same Speaker condition. Furthermore, this Speaker effect on the rate was not modulated by load, supporting the hypothesis that it was due to automatic priming processes.

General discussion

The current article has presented a new solution for analyzing results from eyetracking experiments using a multilevel logistic regression framework. Relative to conventional ANOVA, the logistic regression framework can better accommodate the continuous variable of time, a variable of critical importance in psycholinguistic research. It furthermore can separately estimate anticipatory and rate effects of experimental variables, effects that are confounded when time is collapsed. Additionally, a logistic scale is preferred when the DV is categorical, given that analysis on a proportional scale can yield improper estimation of effects. Finally, the multilevel modeling approach makes it possible to account for nonindependence among observations while minimizing the information loss that can take place due to aggregation.

One aspect of the MLR framework that is in need of further refinement is its handling of eye-movement based dependencies. Currently these are handled either through multilevel logistic regression with robust standard errors (for experiments with many participants and items), or through a less powerful analysis that uses aggregation via the empirical logit (for experiments with a small number of participants or items). These two solutions are effective, but they make it impossible to exploit the full potential of the multilevel approach. First of all, with multilevel approaches it is possible to perform analyses involving ‘crossed’ rather than ‘nested’ random effects (Bates, 2005; Raudenbush & Bryk, 2002), obviating the need for separate participant and item analyses (Baayen et al., this issue). However, robust standard errors are not available for such analyses. Furthermore, the need to aggregate over many trials for empirical logit regression will make use of crossed effects impossible, given that most experiments involve a single presentation of each item to each participant. Ultimately, a better way of handling eye-movement based dependencies would be to develop an algorithm that filters out observations from the data set that are likely to be dependent (e.g., observations recorded during, or proximal to, an eye movement).

Another limitation of the current instantiation of MLR presented here is its ability to handle only a dichotomized dependent variable (i.e., a variable that had only two categories). It cannot handle DVs with more than two categories without first collapsing some of the categories to make the variable binary. However, this is not a limitation of the framework per se, but of the available software. A generalization of logistic regression, known as multinomial logistic regression can be used to handle polytomous variables, and could be readily incorporated into the framework. However, it is presently difficult to fit multinomial logistic regression models using existing software, and robust standard errors may not be provided. Furthermore, it is not clear how to implement an empirical logit transformation with multinomial data. Although an alternative approach using loglinear regression (Knoeferle et al., 2005) can handle a polytomous response variable, this approach, as noted above, has the problem that it cannot accommodate continuous predictor variables; furthermore, it is not clear that it appropriately handles eye-movement based dependencies.

Lastly, it is worth highlighting the generality and flexibility of the MLR approach, which lends itself to research questions beyond those addressed here. Many studies in experimental psychology involve dependent variables that are categorical (e.g., research on language production, where the dependent variable is often the rate that a given linguistic structure is used; studies on recognition memory where the outcome variable is a dichotomous old/new judgment) or an interest in assessing change over time (e.g., experiments on learning where accuracy is measured over a series of blocks). Adoption of the MLR framework will help avoid the pitfalls associated with the analysis of proportional data, as well as provide a more natural framework for analyzing data sets in which time is a critical factor.

Implementation of the MLR framework

This final section provides a more in-depth view of the MLR framework, and is designed for researchers interested in applying it to their own data sets. The reanalysis of Kronmüller and Barr (2007) will be walked through in greater depth. Data files and source code as well as updated information and further examples can be found at the author's website at talklab.org/tvw.

More on logistic regression and the empirical logit

Generalized linear models (GLMs) make it possible to apply the regression framework to data sets with a categorical dependent variable (McCullagh & Nelder, 1989). In a generalized linear model, the relationship between the response variable and the predictor variables is specified via a link function that transforms the response variable onto a linear scale. In a GLM, the error variance can also modeled in terms of distributions other than the normal distribution. For logistic regression, the link function is given by Eq. (3), and residual error is distributed according to a Bernoulli distribution, with variance proportional to the mean p, as given by the formula $p(1-p)$. The logit can be converted back to a probability using the inverse link formula

$$\phi = \frac{1}{1+e^{-\eta}} \tag{5}$$

For multilevel logistic regression, there are two types of estimation procedures that fit two different kinds of models: *unit-specific* (or *conditional*) models and *population-average* (or *marginal*) models. The parameter estimates in a unit-specific model are conditioned on the random effects, while the parameters in a population-average model are not (the random effects are integrated out). The distinction between population-average and unit-specific models is quite confusing because it only arises for multilevel models with a nonlinear link function. One way to grasp the difference is by noting that when aggregating data from a study, one can apply a nonlinear link function at different points. Imagine a two-level model where there is a single random effect for each participant. One can compute a proportion for each participant and then apply the link function on that proportion. The resulting log odds values can then be averaged together. This would be equivalent to a unit-specific model. In contrast, one could average together the proportions from different subjects and then apply the link function on the aggregated proportions; this would be equivalent to a population-average model. These two approaches would yield different outcomes. (For further discussion of these differences, see Fitzmaurice et al., 2004). What is important for the researcher to know is that population-average estimates are based on fewer assumptions, and therefore, are less sensitive to misspecification of the random effects in the model. Therefore, it is recommended that the population-average estimates be used.

When transforming data onto the log odds scale, either for visualization purposes or to filter out eye-movement based dependencies in the data, problems arise applying the link function (3) whenever the probability ϕ approaches zero or one. This is because the resulting η will approach negative or positive infinity. Agresti (2002), and McCullagh and Nelder (1989) recommend instead the empirical logit transformation. To compute the empirical logit, it is necessary to aggregate over multiple observations. Instead of computing proportions and applying the link function, one computes:

$$\eta' = \ln\left(\frac{y+.5}{n-y+.5}\right). \tag{6}$$

In the equation, y is the number of times that the target event was observed, and n is the total number of cases over which y was observed.

When performing empirical logit regression, McCullagh and Nelder (1989) suggest performing a weighted linear regression with weights $1/v$ where

$$v = \frac{1}{y+.5} + \frac{1}{n-y+.5}. \tag{7}$$

Preparing the data

For the analysis using multilevel logistic regression, all individual data points were included in the analysis without aggregation. The time covariates were represented on a scale of seconds rather than milliseconds, otherwise the parameter estimates would have turned out too small (since very little happens in the course of a single millisecond). The time variable was centered at 300 ms, the start of the analysis window, by subtracting .3 s from the timestamp for each frame; for instance, a frame observed 400 ms after the onset of the word would be assigned a value of $.400 - .300 = .100$. The response variable was coded as '1' if the participant's point of gaze was within the target region during this frame, and as '0' otherwise.

For the empirical logit regression, the data were aggregated into a series of four 50 ms bins (three frames of eye data each), starting 300 ms after word onset. Each bin contained the data from the trials in each condition for each participant (or item). For instance, in the participant analysis, each empirical logit was based on 12 frames of eye data (three data frames per trial times four trials per condition). The time codes for each of the four bins were .000, .050, .100, and .150 respectively.

In all eyetracking data sets, frames for which the participant is not gazing at the target region will consist of various cases: the participant is fixating another region; the participant is fixating an empty portion of the screen; the participant is in the midst of an eye blink; or, the eyes are in transit from one location to another. One question is whether to throw out from the data set all frames that do not include a fixation to any viewing region. Using such a procedure, only cases where the participant is fixating a region other than the target would be coded as '0'. As a result, the response variable would represent the log odds of fixating the region of interest, given that the participant is fixating some viewing region.

However, this scheme assumes that the 'noise' events are equally distributed across conditions, which may not be the case if, for instance, participants are slower to move off of a central fixation cross in one condition than in another, or if their blink rate varies across conditions (perhaps due to differing attentional demands). Therefore, it is safer to code all frames including these 'noise' frames as '0' instead of throwing them out of the data set.

Time-locking and determining the form of the model

The first decisions that must be made concern the time-locking the analysis window, as well as the form of the model that will be fit to the data. Ideally, one should time-lock the analysis window at the earliest point at which gaze behavior driven by the linguistic signal emerges. As noted previously, gaze behavior that is driven by language processing will typically be delayed by a latency of 180–200 ms, because of the time needed to program an eye movement (Matin et al., 1993). Furthermore, in the current experiment, the signal-driven eye movements are likely to have emerged even later than in a other eyetracking experiments, since unconventional descriptions were used to designate unfamiliar pictures. Thus, the ability to differentiate a target object would require more linguistic evidence than in other experiments when objects with conventional names are used.

The point at which signal-driven eye movements appear in the eye data will generally appear as a sudden rise in fixations toward a target object. The data prior to this point will generally not be of interest to the researcher. Furthermore, attempting to include the data prior to this point can either lead to a poor fit or require a more complex model than the data demand, thus needlessly complicating the interpretation of the parameters of the model.

One way of determining the appropriate time-locking is to plot the grand mean data and time-lock the data to the first frame that begins a rising trend. This is a conservative procedure: by time-locking based on the grand mean, one can avoid biasing one's choice toward any given hypothesis.

The next important choice is the form of the model that will be fit to the data. Of course, the form of the model will ultimately depend on the researcher's goals as well as precedents from previous research, of which there are currently very few (see for instance, Scheepers et al., in press; Magnuson et al., 2007). In the end, it is best to ground such choices in underlying theory to the extent possible. Sometimes, however, choices must be made by examining the observed response curves. Under such circumstances, there are several considerations to keep in mind. First, the curves that are being modeled in logistic regression are on a log odds scale, not on the proportional scale. Thus, it is better to make judgments about the appropriate form of the model while viewing data on the log odds scale. As shown in Fig. 2, curvature that appears on a proportional scale may disappear when plotted on a log odds scale, and vice versa. In visual world experiments, response functions often have a 'S' or sigmoidal shape when plotted on a proportional scale. However, when the same data are plotted on a log odds scale, the bends in the function will tend to straighten out, such that the data could be parsimoniously fit by a strictly linear function.

By counting the number of bends in a response curve, one can get a sense of the order of the equation that is necessary to fit the data. It is useful to start with an equation of sufficiently high order to capture the trends for the condition with the greatest number of bends. For example, with two bends, a cubic equation might be required; with only a single bend, a quadratic will be sufficient. Model selection techniques can be used to assess whether higher-order terms can safely be excluded from the model. It is important to note, as mentioned above, that inclusion of higher-order terms can change the interpretation of lower order terms in the model (Cohen et al., 2003). For instance, if a quadratic term is included, then the term corresponding to the slope will index the slope of the line tangent to the curve at $t = 0$ (i.e., the instantaneous slope at the intercept) rather than the slope throughout the window.

A further cautionary note is that for fourth order or higher equations, the parameters can be extremely difficult to interpret. Under such circumstances, an alternative method is to individually model smaller segments of the line, either by conducting separate analyses or by dividing up the curve into polynomial "splines". For further information on splines, see Fitzmaurice et al. (2004) and Snijders and Bosker (1999).

The multilevel model

Level-1 of the multilevel model expressed the log odds η of a '1' response for frame *f*, trial *i* and participant *p* as a function of time using the following equation.

$$\eta_{fip} = \pi_0 + \pi_1 t_{fip} \tag{8}$$

At level-2 (the trial level), the level-1 coefficients were modeled in turn as a function of fixed and random effects associated with the given trial *i* for participant *p*. In the empirical logit analysis, observations from trials in a given condition were collapsed; thus, level-2 was defined by condition in the experiment, rather than by trial.

At level-2 the variable codes for the main effect of Speaker and Load were entered, as well as for an interaction effect. Because Speaker and Load are categorical variables, an effects coding scheme (Cohen et al., 2003) was used so that the parameter estimates would correspond to the tests for main effects in a standard ANOVA; i.e., tests on marginal means. Note that given the factorial design, a dummy coding scheme for the same variables (that is, using 0 and 1) would result in each variable representing the simple effect of the variable in the condition of the other variable coded as '0'; i.e., a test on cell means rather than on marginal means.

In the effect coding scheme, the values of $-.5$ and .5 were chosen so that the parameter estimates would be equivalent to the mean difference in log odds between the two conditions. The interaction term was calculated by multiplying the signs (i.e., positive or negative) of the two Speaker and Load effect codes together and then multiplying that result by .5. The interaction therefore reflects a contrast between the two sets of cells lying across opposite diagonals in the design matrix. The test of the significance of this parameter estimate is directly equivalent to the standard ANOVA test for an interaction.

Equations for the model at level-2 are given below.

$$\begin{aligned}\pi_0 &= \beta_{00} + \beta_{01}S_{ip} + \beta_{02}L_{ip} + \beta_{03}SL_{ip} + r_0\\ \pi_1 &= \beta_{10} + \beta_{11}S_{ip} + \beta_{12}L_{ip} + \beta_{13}SL_{ip} + r_1\end{aligned} \tag{9}$$

The first subscript of each of the level-2 βs corresponds to the subscript of the level-1 term π on the left side of the equation (0 for intercept and 1 for slope), while the second subscript indexes the role of the coefficient in the level-2 equation. The random effects r_0 and r_1 allow the intercepts and slopes (respectively) to vary randomly across trials (or across conditions, for the empirical logit analysis). Note that the subscript used for the random effects is the same as that of the level-1 coefficient that is being predicted.

Given the effect coding scheme, β_{00} and β_{10} will correspond to the estimated grand mean for the intercept and the estimated grand mean for the slope, respectively. In regression, significance tests of parameter estimates test the null hypothesis that $\beta = 0$. The effect coding scheme is useful because the significance tests for the coefficients of the predictor variables corresponds to the significance tests for main effects in an ANOVA. Thus, the test for β_{01} and β_{11} corresponds to the test for the main effect of Speaker on the intercept and slope, respectively. Similarly, the test for β_{02} and β_{12} correspond to the test for the main effect of Load on the intercept and slope; and the test for β_{03} and β_{13} correspond to the test for an interaction effect.

At level-3 of the MLR framework the level-2 β coefficients are modeled in terms of level-3 fixed and random effects. If there are any measured characteristics of the relevant level-3 units, they can be included at this level. It is also possible to test for cross-level interactions between these measured covariates and the level-2 IVs. Mostly, however, level-3 will include only random effects, in order to account for the clustering of observations within participants (or items).

One decision that must be made is where to include random effects at level-3. Note that there will be a separate equation at level-3 for each of the βs in the level-2 equations. For the current data set, then, one could in principle include up to eight random effects at level 3, one for each β in Equation Set (9). In practice, however, this would be impractical because the estimation algorithms used for fitting multilevel models are iterative and become less likely to reach a solution as the number of random effects increases.

Including a random intercept for β_{00} assumes that the intercept term in Eq. (8) varies over participants. Recall that the intercept corresponds to the likelihood of looking at the region coded as '1' at the very start of the stimulus presentation. It is likely that this would vary over subjects, with some following different anticipatory strategies than others. Likewise, including a random intercept for β_{10} assumes that the slope term in Eq. (8) varies over participants. This also would seem to be a valid assumption.

In contrast, including a random intercept for β_{01} would assume that the effect of X_{ip} on the intercept varies across participants. Now, the idea that the magnitude of the effect of an IV can vary across participants is not one that most researchers typically consider; but this is largely because it is not possible to model such effects within the ANOVA framework. In some cases, it may be reasonable to assume that certain participants will differ in their susceptibility to a manipulation than others. For example, participants with a large working memory span may show smaller effects of a memory load manipulation than participants with a smaller span. But unless the goal of the research is to examine such effects, parsimony would suggest leaving these terms out.

Therefore, for the current example we include random effects only for β_{01} and β_{10}. Given this, we can simplify our equations by inserting the level-3 random effects into our level-2 equations, rather than stating another series of equations. Thus, the level-2 equations become

$$\begin{aligned}\pi_0 &= \beta_{00} + \beta_{01}S_{ip} + \beta_{02}L_{ip} + \beta_{03}SL_{ip} + r_0 + u_0,\\ \pi_1 &= \beta_{10} + \beta_{11}S_{ip} + \beta_{12}L_{ip} + \beta_{13}SL_{ip} + r_1 + u_1,\end{aligned} \tag{10}$$

where the u variables represent random participant (or item) effects.

It is worth noting that the specification of equations at multiple levels is simply a notational convenience that is meant to capture the different levels of sampling in the model. The entire model can ultimately be collapsed and expressed a single equation, called the "mixed model", by inserting the level-3 equations into level-2, and then the resulting equations into level-1. Indeed, the algorithms for parameter estimation ultimately estimate the level-3 coefficients from the mixed model. The output of these algorithms provides significance tests only for these γ coefficients.

Once parameter estimates are obtained, one can plot the model for the response curves by substituting the parameter estimates for the fixed effects into the equations in set (10), generating values for π_0 and π_1 for each of the four conditions, and then substituting these values in at level-1 to create one equation for each condition. The resulting equations can also be used to generate a model on the probability scale using the inverse link function (5).

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: Wiley.

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73*, 247–264.

Arai, M., van Gompel, R. P., & Scheepers, C. (2007). Priming ditransitive structures in comprehension. *Cognitive Psychology, 54*, 218–250.

Baayen, R. H. (2007). languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics". R package version 0.2.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (submitted for publication). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*.

Barr, D. J., & Keysar, B. (2002). Anchoring comprehension in linguistic precedents. *Journal of Memory and Language, 46*, 391–418.

Bates, D. (2005). Fitting linear mixed models in R. *R News, 5*, 27–30.

Bates, D. (2007). *lme4: Linear mixed-effects models using S4 classes*. (R package version 0.99875-2).

Church, B. A., & Schacter, D. L. (1994). Perceptual specificity of auditory priming: Implicit memory for voice intonation and fundamental frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 521–533.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Mahwah, NJ: Erlbaum.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology, 6*, 84–107.

Dahan, D., & Tanenhaus, M. K. (2004). Continuous mapping from sound to meaning in spoken-language comprehension: Immediate effects of verb-based thematic constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 498–513.

R Development Core Team. (2007). R: A *language and environment for statistical computing.* Vienna, Austria. (ISBN 3-900051-07-0).

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis*. New York: Wiley.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 1166–1183.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). Thousand Oaks, CA: Sage.

Hanna, J., Tanenhaus, M. K., & Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language, 49*, 43–61.

Hoffman, L., & Rovine, M. J. (2007). Multilevel models for the experimental psychologist: Foundations and illustrative examples. *Behavior Research Methods, 39*, 101–117.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). Hoboken, NJ: Wiley.

Howell, D. C. (1997). *Statistical methods for psychologists* (4th ed.). Belmont, CA: Wadsworth.

Jaeger, T. F. (submitted for publication). Categorical data analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models. *Journal of Memory and Language*.

Knoeferle, P., Crocker, M. W., Scheepers, C., & Pickering, M. J. (2005). The influence of the immediate visual context on incremental thematic role-assignment: Evidence from eye-movements in depicted events. *Cognition, 95*, 95–127.

Kronmüller, E., & Barr, D. J. (2007). Perspective-free pragmatics: Broken precedents and the recovery-from-preemption hypothesis. *Journal of Memory and Language, 56*, 436–455.

Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., & Aslin, R. N. (2007). The dynamics of lexical competition during spoken word recognition. *Cognitive Science, 31*, 1–24.

Matin, E., Shao, K. C., & Boff, K. R. (1993). Saccadic overhead: Information processing time with and without saccades. *Perception & Psychophysics, 53*, 372–380.

McCullagh, P., & Nelder, J. (1989). *Generalized linear models*. London: Chapman and Hall.

Metzing, C., & Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language, 49*, 201–213.

Nadig, A. S., & Sedivy, J. C. (2002). Evidence of perspective-taking constraints on children's on-line reference resolution. *Psychological Science, 13*, 329–336.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics, 60*, 355–376.

Quené, H., & van den Bergh, H. (2004). On multi-level modeling of data from repeated measures designs: A tutorial. *Speech Communication, 43*, 103–121.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.

Richter, T. (2006). What is wrong with ANOVA and multiple regression? analyzing sentence reading times with hierarchical linear models. *Discourse Processes, 41*, 221–250.

Scheepers, C., Keller, F., & Lapata, M. (in press). Evidence for serial coercion: A time course analysis using the visual-world paradigm. *Cognitive Psychology*.

Snijders, T. A., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

Tanenhaus, M. K., Spivey, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*, 1632–1634.

Please cite this article in press as: Barr, D. J., Analyzing ‘visual world’ eyetracking data using ..., *Journal of Memory and Language* (2008), doi:10.1016/j.jml.2007.09.002