PSYCHOMETRIKA—VOL. 85, NO. 1, 154–184
MARCH 2020
https://doi.org/10.1007/s11336-020-09694-6

MODELING INTENSIVE POLYTOMOUS TIME-SERIES EYE-TRACKING DATA: A DYNAMIC TREE-BASED ITEM RESPONSE MODEL

SUN-JOO CHO

VANDERBILT UNIVERSITY

SARAH BROWN-SCHMIDT

VANDERBILT UNIVERSITY

PAUL DE BOECK

THE OHIO STATE UNIVERSITY

KU LEUVEN

JIANHONG SHEN

VANDERBILT UNIVERSITY

This paper presents a dynamic tree-based item response (IRTree) model as a novel extension of the autoregressive generalized linear mixed effect model (dynamic GLMM). We illustrate the unique utility of the dynamic IRTree model in its capability of modeling differentiated processes indicated by intensive polytomous time-series eye-tracking data. The dynamic IRTree was inspired by but is distinct from the dynamic GLMM which was previously presented by Cho, Brown-Schmidt, and Lee (Psychometrika 83(3):751–771, 2018). Unlike the dynamic IRTree, the dynamic GLMM is suitable for modeling intensive binary time-series eye-tracking data to identify visual attention to a single interest area over all other possible fixation locations. The dynamic IRTree model is a general modeling framework which can be used to model change processes (trend and autocorrelation) and which allows for decomposing data into various sources of heterogeneity. The dynamic IRTree model was illustrated using an experimental study that employed the visual-world eye-tracking technique. The results of a simulation study showed that parameter recovery of the model was satisfactory and that ignoring trend and autoregressive effects resulted in biased estimates of experimental condition effects in the same conditions found in the empirical study.

Key words: autocorrelation, eye-tracking data, generalized linear mixed effect model, intensive polytomous time series, multinomial processing tree, tree-based item response model, trend.

1. Introduction

1.1. Nominal Eye-Tracking Data

Eye-tracking systems are inexpensive and widely available, and eye movements can be used as a time-sensitive measure of multiple cognitive processes. In recent years, eye tracking has been used in a wide variety of disciplines including psychology, reading education, medical research (neurological diagnosis), and marketing (see Richardson & Spivey, 2004 for a review). Eye-tracking systems are capable of automatically generating fixation location data over time.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11336-020-09694-6) contains supplementary material, which is available to authorized users.

Correspondence should be made to Sun-Joo Cho, Vanderbilt University, Nashville, USA. Email: sj.cho@vanderbilt.edu; URL: http://www.vanderbilt.edu/psychological_sciences/bio/sun-joo-cho

The location data can be nominally coded based on which delineated area of a viewed surface (i.e., a cell of a grid) the viewer looks at each time point. Depending on the number of nominal categories of interest, the fixation data can be coded as binary or polytomous time-series data. The analysis of the binary time-series data aims to identify and analyze visual attention to a single interest area over all other possible fixation locations (e.g., Cho, Brown-Schmidt, & Lee, 2018). In contrast, the analysis of polytomous time-series data aims to understand the visual attention given to several competing options possibly associated with different processes. The motivation for the present work is a research question from the field of psycholinguistics for which multiple cognitive processes are assumed to differentially map onto one or more of several competing response options. To the best of our knowledge, this is the first attempt to model the multiple cognitive processes in polytomous eye-tracking data in the literature.

1.2. Multinomial Processing in Eye-Tracking Data

In complex scenes, eye gaze is probabilistically directed to individual fixation locations, with the likelihood of a fixation to any particular location driven by several competing or complementary cognitive processes (e.g., McMurray, Samelson, Lee, & Tomblin, 2010; McMurray, Klein-Packard, & Tomblin, 2019; Mozuraitis, Chambers, & Daneman, 2015). In cases where eye gaze is in service of performing a task, one of the locations can be considered a task-relevant "target" location (e.g., an object that a person will select), another location may be similar to the target on some dimension, resulting in potential confusion between the target and this "competitor" object. Lastly other locations may be unrelated to the target and less likely to receive visual attention. In such situations, we expect that multinomial processing will guide the likelihood of fixating each of the response categories, with one cognitive process increasing the likelihood of fixations to the target and competitor, and a separate cognitive process that allows the viewer to select the target and rule out the competitor.

1.3. Generalized Linear Mixed Effect Model (GLMM) and Dynamic GLMM

In the statistics literature, model specification for time-series data has primarily employed generalized linear models (GLMs) and the resulting models have been termed dynamic GLMs, or exponential family state-space models (e.g., Fahrmeir, 1992; Gamerman, 1998; West, Harrison, & Migon, 1985). The generalized linear mixed effect model (GLMM) permits the response probability distribution to be any member of the exponential family of distributions (e.g., Binomial, Poisson, Gamma) and provides a flexible modeling framework to model heterogeneity and dependency between observations. A GLMM can be called a dynamic GLMM when change processes such as trend and/or serial autocorrelation in time-series data are considered. Hung, Zarnitsyna, Zhang, Zhu, and Wu (2008) specified a dynamic GLMM for binary time-series data in which random effects were employed to model heterogeneity among experimental units (e.g., participants). In existing dynamic GLM or GLMM specifications, heterogeneity among items is often ignored. When items are heterogeneous, ignoring such heterogeneity in a GLMM can lead to biased parameter estimates and underestimated standard errors of fixed effects in the GLMM (e.g., Baayen, Davidson, & Bates, 2008; Jaeger, 2008; Quené & van den Bergh, 2008). In the psychometrics literature, existing models incorporate latent variable-based serial dependence (e.g., Browne & Nesselroade, 2005; Molenaar, 1985). However, Cho et al. (2018) noted that the literature lacks presentation of a model that simultaneously models random item effects and heterogeneity in autoregressive (AR) parameters across persons and items for binary time-series data. To overcome this limitation of existing models, Cho et al. (2018) presented a dynamic GLMM in which heterogeneity in AR parameters across persons and items was modeled for binary (e.g., target vs. non-target fixation in eye-tracking data) time-series data. Cho et al. (2018) focused on visual attention to a single interest area (the target) in detecting experimental condition effects (controlling for AR). The dynamic GLMM is a model for binary data and binomial processes so that it cannot accommodate theories which stipulate differential processing depending on the response option.

1.4. A Response Tree Approach and Dynamic IRTree Model

Tree-based item response (IRTree) models (De Boeck & Partchev, 2012) are tree models for polytomous data, which implies that each response option is reached through a unique branch path in a tree. Tree models are inspired by sequential item response models (Tutz, 1990) and multinomial processing tree models (Riefer & Batchelder, 1988; see Batchelder & Riefer, 1999 for a review). The hypothesized processes at each node in the tree can be interpreted as sequential, but strictly speaking, the tree only implies that the choices at the lower nodes in the tree are *conditional* choices. For example, the tree we will work with has a top node A vs. B (i.e., unrelated objects vs. target & competitor objects) and a lower-order node B1 vs. B2 (i.e., target vs. competitor), so that B1 and B2 are conditional choices given B is chosen at the top node. A tree model is an IRTree model if the choices at the nodes in the tree are modeled with an item response model. The item response model can be of any kind (for an overview of item response models, see, for example, Embretson & Reise, 2000), including a GLMM as is commonly the case in psycholinguistics (e.g., Baayen et al., 2008). For an overview of the possibilities of IRTree models, see Böckenholt (2017) and Jeon and De Boeck (2016). While the original multinomial processing tree models did not allow for person and item heterogeneity, the models have later been extended to accommodate such heterogeneity (Batchelder & Crowther, 1997; Böckenholt, 2012; Klauer, 2010; Matzke, Dolan, Batchelder, & Wagenmakers, 2015; Walker, Hickok, & Fridriksson, 2018). Because the sequential item response models are item response models (Tutz, 1990), already the original formulation of the models did allow for the heterogeneity.

However, neither IRTree models nor multinomial processing tree models have been extended to allow for trend and AR effects. Literature reviews on existing IRTree models and time-series models lead to the conclusion that there is a disconnection between the available analytic methods and a common data structure in studies of real-time cognitive processes using the visual-world eye-tracking technique (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). This disconnection can be resolved by combining the IRTree model (De Boeck & Partchev, 2012) with the time-series model (e.g., see Chatfield, 2004, for a review). The goal of the novel modeling framework in the present study is to allow for: (a) differential processing depending on the response option (based on the tree feature of the model), (b) heterogeneity of the processes (based on the item response theory [IRT] feature of the model), and (c) change processes (trend and AR parameters) as in the time series models (e.g., Craigmile, Peruggia, & Van Zandt, 2010) and in the dynamic GLMM. The difference between the present approach and the dynamic GLMM is that the time series is modeled separately by node in the tree and thus possibly different trend and AR effects depending on the node. For each node, the dynamic effects (trend and AR) can be heterogeneous, varying across persons and items. The model is called a *dynamic IRTree model*. Like for the dynamic GLMM, ignoring trend and AR and their heterogeneity can lead to bias in the estimates of interest as we will show. It has been found that conclusions about the AR effect can be inaccurate in the presence of unmodeled trend effects (e.g., Jahng, Wood, & Trull, 2008; Wang, Hamaker, & Bergeman, 2012), and we will show that modeling trend and AR and their heterogeneity are important to avoid bias in the estimates of interest.

The novelty of the dynamic IRTree model lies in the combination of three features: the tree feature, the IRT feature, and the dynamic feature. All three are important to answer substantive research questions regarding cognitive processes underlying data from a linguistically inspired eye-tracking study. Furthermore, as we explain in a later section, applying a tree-based model to the data results in an extra challenge, which is the problem of how to deal with necessarily missing observations regarding the lower node. If a person fixates an unrelated object (the A choice), the B1 vs. B2 observation (target or competitor) is necessarily missing. Thus, we show how to model the missing observations for AR effects in this study, which has yet to be demonstrated in the literature.

The remainder of this paper is organized as follows: In Sect. 2, we describe an empirical study. In Sect. 3, we present the dynamic IRTree model, provide a parameter estimation method, and describe the model selection, testing, and evaluation methods. In Sect. 4, the dynamic IRTree model is illustrated using an empirical data set. In Sect. 5, parameter recovery of the model and consequences of various misspecifications in detecting experimental condition effects are evaluated via a simulation study. In Sect. 6, we end with a summary and a discussion.

2. Motivating Data: Intensive Polytomous Time-Series Eye-Tracking Data

We apply a dynamic IRTree model to an empirical data set to illustrate the use of the model and to give motivation for its development. In this section, we describe the data set that multinomial processing is assumed to guide. The data set comes from a study previously published as Ryskin, Benjamin, Tullis, and Brown-Schmidt (2015). Cho et al. (2018) analyzed the raw data in binary form (i.e., whether the participant was looking at a target); they did not attempt a multinomial processing approach using polytomous data (i.e., distinguishing whether the participant was looking at a target, a competitor, or unrelated objects).

2.1. Eye-Tracking Data

One hundred and fifty-two native English-speaking participants from the University of Illinois at Urbana-Champaign student community participated in exchange for partial course credit. The participants completed the study in pairs. The participants were seated at separate computers, each of which was equipped with an EyeLink 1000 eye-tracker (SR Research). The focus of the analysis is on the eye-tracking data that were collected as the participants took turns instructing each other to click on objects on the computer screen. The eye-tracking data provide an online measure of the millisecond-by-millisecond cognitive processes by which the listener interprets phrases like “the small elephant,” “the banana,” “the large dog,” etc. Each experimental screen contained a 3 × 3 grid with pictures, some of which were seen by both speaker and listener (white background), and some of which were only seen by one person. The latter type of picture was shown on a gray background to the person who could see it and a black square to the other person (see Fig. 1, left panel). This manipulation of who can see what provides the basis for measuring perspective-taking and is based on earlier work which featured real objects in a real display, some of which were visible to only one person (e.g., Nadig & Sedivy, 2002). Subsequently, the computer-based adaptation of the task has been used with success with both healthy adults (e.g., Brown-Schmidt, Gunlogson, & Tanenhaus, 2008; Brown-Schmidt, 2009a; 2009b; 2012) and memory-impaired individuals (Rubin, Brown-Schmidt, Duff, Tranel, & Cohen, 2011). The right side of Fig. 1 illustrates an example of polytomous eye-tracking data from a single trial in the data set. The top right panel plots the data over a 6-second period of time, with the polytomous responses denoted by color. The x-axis is in milliseconds, beginning at trial onset; the y-axis corresponds to the original fixation position x-coordinate of the display computer. The bottom right panel plots the 3 fixation categories over time (on the x-axis) from 180 to 1300 milliseconds following the onset of the critical word, e.g., “small” in the phrase “the small elephant.” The y-axis denotes fixations to the target (small elephant), the competitor (small envelope), and everything else (note this category includes fixations to all other interest areas on the screen, as well as time points with no fixations).

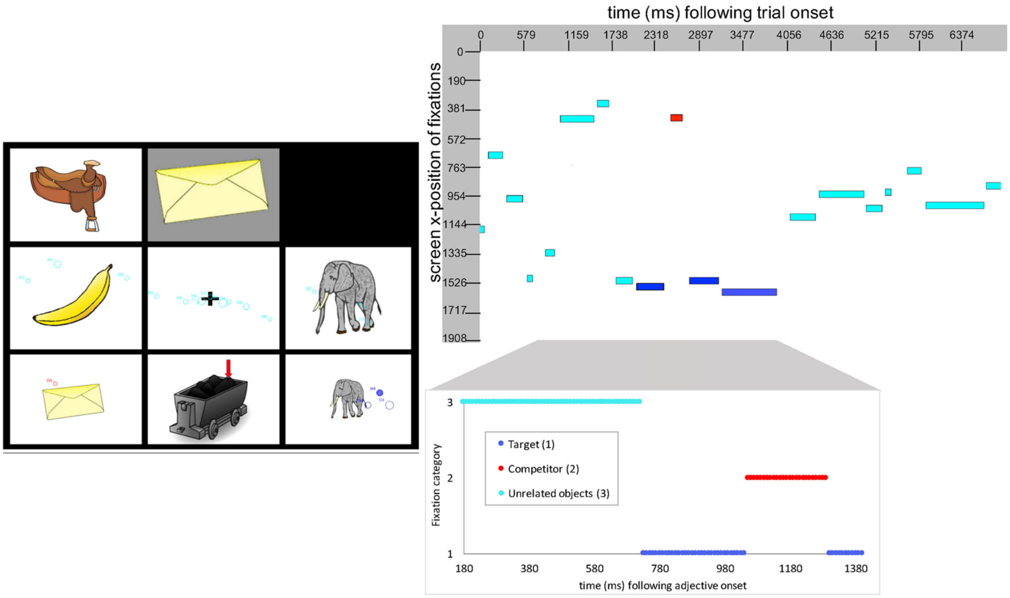

FIGURE 1.
Illustration of the collection and processing of the eye-tracking data for a single trial. The left panel is an example of visual stimulus from the perspective of one participant; their partner would view a similar scene. The images in white were visible to both participants. The images in a gray background were visible to one participant. (The other participant saw a black box in this spot instead.) Participants received instructions about which images they could both see (shared perspective), and which images only they could see (non-shared perspective); this afforded the critical manipulation of visual perspective in the experiment. Across trials, the images and their positions in the 3 × 3 grid were varied following standard experimental controls. Superimposed on the visual stimulus are circles corresponding to individual fixations (dark blue = target; red = competitor; light blue = unrelated objects). The top right panel shows the sequence of fixations over time, from trial onset on the x-axis, with fixations visually separated by their x-position on the visual screen (in pixels) shown on the y-axis. Each fixation is color-coded as before (dark blue = target; red = competitor; light blue = unrelated objects). The bottom right panel zooms in on the time region of interest which begins 180 milliseconds after critical adjective onset and illustrates the polytomous nature of the data with the participant on this trial looking at an "other" unrelated object, then the target, the competitor and back to the target at the very end.

2.2. *Experimental Design*

Three experimental conditions were manipulated with a within-subjects design. In the *Two Contrasts-Shared* condition, the speaker instructed the listener to, for example, "Click on the small elephant." The listener's screen showed the target referent (small elephant), an item of the same type as the target that contrasted in size (large elephant), a competitor that was the same size as the target and had a similar sounding name (small envelope), an item contrasting the competitor in size (large envelope), and three unrelated items (Fig. 1, left panel). The *Two Contrasts-Privileged* condition were similar in that the listener saw the same scene, with the exception that the large envelope (the size contrast for the competitor) was visible to the listener but not to the speaker. To illustrate this perspective difference, the size contrast appeared with a gray background to the listener, and on the speaker's screen this picture was replaced by a black square. The experimental instructions made clear that objects in gray backgrounds were seen only by one partner. As the focus of this research is on perspective-taking, this perspective manipulation is the key experimental manipulation. Finally, in the *One-Contrast* condition, the item that contrasted the competitor in size (e.g., the large envelope) was replaced by an unrelated item (e.g., a train). Standard experimental controls including randomization of trial order, items, and picture positions were employed. The empirical study was designed to examine the influence of contextual and pragmatic factors on the time course of eye fixations as the listener interprets the critical expression, e.g., "small elephant." Based on prior work examining spoken language processing, it is assumed that the processing of the initial speech sounds results in an activation of candidate referents which results in fixations to those candidate referents. Ambiguity between candidate referents can be resolved through various sources of information including subsequent words and phrases as well as contextual factors, allowing the listener to identify and fixate the target (e.g., Allopenna, Magnuson, & Tanenhaus, 1998; Hanna, Tanenhaus, & Trueswell, 2003; Tanenhaus et al., 1995).

Upon hearing the unfolding expression, the lexico-semantic information (the meaning of the words "the small e-") is consistent with the target and the competitor, but not with the unrelated objects. The focus of this experiment is on the fixed condition effects. Figure 1 illustrates an example trial from the Two Contrasts-Privileged condition, where the contrasting object for the competitor (large envelope) is in a gray background and thus visible to the listener but not to the speaker. In this condition, pragmatic information (i.e., how words are used in context) differentiates the target from the competitor because, from the speaker's perspective, only the small elephant is the smaller of two items from the same category. While the competitor (small envelope) temporarily matches the lexico-semantic information ("the small e-"), it is not consistent with the perspective information in this condition (i.e., the speaker would have called it "the envelope" rather than "the small envelope" because the speaker only sees one envelope).

2.3. Multinomial Processing

The motivating example concerns listeners' interpretation of instructions, e.g., "Click on the small elephant" in scenes that contained pictures of seven objects, including a small elephant (the target), a small envelope (the competitor), and five other unrelated objects that were not candidate referents (see Fig. 1, left panel). It is assumed that the likelihood of fixating each of the three object categories is guided by multinomial processing: lexico-semantic processing narrows the set of candidate referents to the target and competitor (e.g., small elephant and small envelope). Then, ambiguity resolution processes narrow down the search space, picking out the target (small elephant) over the competitor (small envelope) in one of the experimental conditions. Lexico-semantic information concerns the meaning of words, and in this data set this information differentiates the target and the competitor vs. unrelated objects. The ambiguity between target and competitor can be resolved using different sources of information, including, among other things, information about the perspective of the speaker. To model these multinomial processes, we use a nested design with nested contrasts. The first node in the tree distinguishes objects that match the lexico-semantic information in the unfolding expression vs. those that do not (e.g., small elephant & small envelope vs. everything else). Among the items that match the unfolding expression, the second node in the tree distinguishes the target object from the competitor object (e.g., small elephant vs. small envelope). The dynamic IRTree approach allows us to disentangle complex relationships among different cognitive processes and different factors of interest. For example, it is possible that a given factor has an effect only on the first node of the tree (lexico-semantic processing), but not on the second node (ambiguity resolution), or vice versa. As we will see, separate consideration of the distinct cognitive processes involved is made possible by a response tree approach, leading to new and more differentiated findings.

2.4. Data Structure

One hundred and fifty-two participants completed the experiment in pairs (76 pairs total). Each pair of participants completed a total of 288 trials together (each person also separately completed a number of individual differences measures of, e.g., working memory capacity, which were used as person-level covariates in the original analysis; see Ryskin et al., 2015). On each of the 288 trials, each participant either said something to their partner (e.g., “Click on the big duck”) or heard their partner say something to them (e.g., “Click on the small elephant”). Of the 288 trials, 96 are fillers and not of interest for analysis. Ninety-six are of interest when analyzing the performance of one partner (let us call them “partner 1”), and the remaining 96 trials are of interest when analyzing the performance of the other partner (“partner 2”). Of the trials of interest for analyzing the performance of a given partner, about 1/2 are measuring their language *production* abilities (in the role of speaker), and 1/2 measure their language *comprehension* (in the role of listener). Whereas the original paper reported on both production and comprehension, it is only the comprehension trials that are of interest to the present analysis. Partner 1’s comprehension trials were in one of 3 conditions (Two Contrasts-Shared; Two Contrasts-Privileged; One Contrast). Likewise, Partner 2’s comprehension trials were in one of three conditions (Two Contrasts-Shared; Two Contrasts-Privileged; One Contrast). Each trial featured an “item,” where the item was the object they had to click on (e.g., duck, frog, elephant). There are 96 unique items total in the dataset. Following standard assumptions in the field of psycholinguistics, the items were assumed to be sampled from the population of all items (Clark, 1973). Across all 288 trials, items were repeated multiple times (3 times each), sometimes in the same condition, sometimes in another condition, sometimes on filler trials. Due to the use of a naturalistic paradigm in which a genuine participant produces the critical instructions live, trials were occasionally lost when the speaker made a speech error (e.g., “the elephant oh the small one”). The final data set which we examine here is structured as follows: the total number of all trials in the entire data set is 43,776 ($= 152 \times 96 \times 3$), including filler trials, trials of interest to the other participant in the pair, and the production and comprehension trials for the current participant. Of the 14,592 ($= 152 \times 96$) trials of interest for a given participant, 8886 contain language comprehension time-series data in the conditions of interest. The remaining 5706 cells ($= 14,592 - 8886$) are not considered for analyses because they were excluded from analysis for reasons described above.

The eye-tracking data collected as the listener interpreted the critical instructions on these 8886 trials forms the basis of our analyses. Each of these trials yields eye-fixation data for 112 equally spaced time points (180–1300 milliseconds following the onset of the critical adjective small/large at each of a series of 10 millisecond time bins; see below for discussion of time-window selection). Non-fixation events (blinks and saccades) were treated in the same way, attributing the duration of blink or saccade to the next object that was fixated; as a result, there were no missing polytomous data in the time series for a given trial, prior to modeling multinomial processing.

The data have a multilevel structure, as presented in Fig. 2. The time-series eye-fixation data define Level 1. These time-series data are nested in 288 trials, which define Level 2. The trials themselves are nested by persons and items, which are crossed and define Level 3. As noted in Cho et al. (2018), a preliminary analysis indicated the clustering due to pairs can be ignored.

3. Modeling Multinomial Processing in Polytomous Time-Series Eye-Tracking Data

In this section, we present (a) a dynamic IRTree model, (b) its estimation, and (c) model selection, testing, and evaluation methods.

3.1. A Dynamic IRTree Model

3.1.1. Response Tree Coding of Polytomous Time-Series Data To specify a dynamic IRTree model, the original responses are recoded into binary responses. The polytomous responses are coded as nested data (De Boeck & Partchev, 2012). Figure 3 presents the tree diagram for a three-category paradigm. Each node has binary outcomes. In the figure, circles represent nodes, arrows represent branches, and rectangles represent polytomous responses, y_{tlji}s for time t, trial l, person

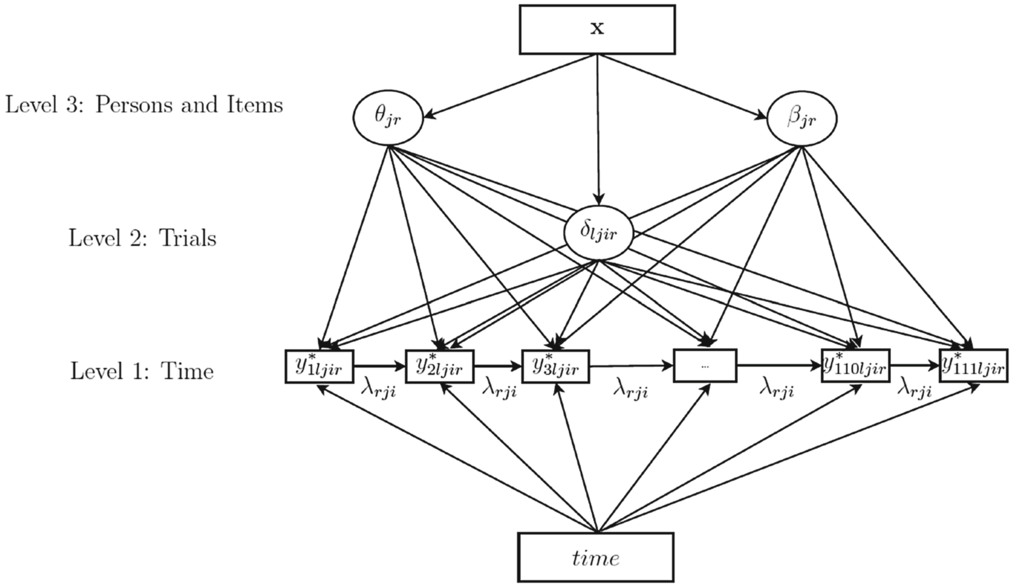

FIGURE 2.
A diagram for the dynamic IRTree model (Eq. 1).

j, and item i. Each node of the tree represents an hypothesized process, and the branches represent the outcomes. They are either intermediate outcomes leading to the next node, or they lead to final response options, also called terminal nodes or leaves, y_{tlji}. We denote the recoded response for the first intermediate node (hereafter, called "Node 1") by y^*_{tlji1} and the recoded response for the second intermediate node (hereafter, called "Node 2") by y^*_{tlji2}. As shown in Fig. 3, one branch from Node 1 leads directly to a terminal node (response category $y_{tlji} = 3$ corresponding to all "other" non-target and non-competitor objects), whereas the other branch leads to Node 2. The branches from Node 1 capture the first differentiation, (a) target and competitor vs. unrelated objects. The branches from Node 2 capture the second differentiation, (b) target vs. competitor. As summarized in Table 1 (top), the recoded binary responses are denoted as $y^*_{tljir} = NA, 0,$ or 1, with $r = 1, 2$ as an index for the node. In the binary data analysis presented in Cho et al. (2018), multinomial processing in the gray area of Fig. 3 was not modeled.

3.1.2. Missing Observations There are missing observations at Node 2. The missingness (NA) at Node 2 results from the structure of the tree. Specifically, the responses (target vs. competitor) at Node 2 of the tree structure (see Fig. 3) are *conditional* responses. The missingness is missingness at random (MAR) according to the definition of MAR (Rubin, 1976),[1] so that it does not affect model parameter estimation. However, when AR parameters are included in the model, it creates a problem for the Node 2 observations at time point t that do not have a corresponding observation at time point $t - 1$. Table 2 illustrates the covariates for the $AR(1)$ effects over the first ten time points for a trial, a person, and an item (top), the covariates at Node 1 (middle), and the covariates at Node 2 (bottom). Because the missingness due to $y_{tlji} = 3$ differs depending on the combination of trial l, person j, and item i over time in Node 2, the number of time points can be different across those combinations at Node 2 (unbalanced design; median = 31 and interquartile

[1] Missing at random is missingness given other observations which means that all the information about when missingness occurs is contained in the observed information.

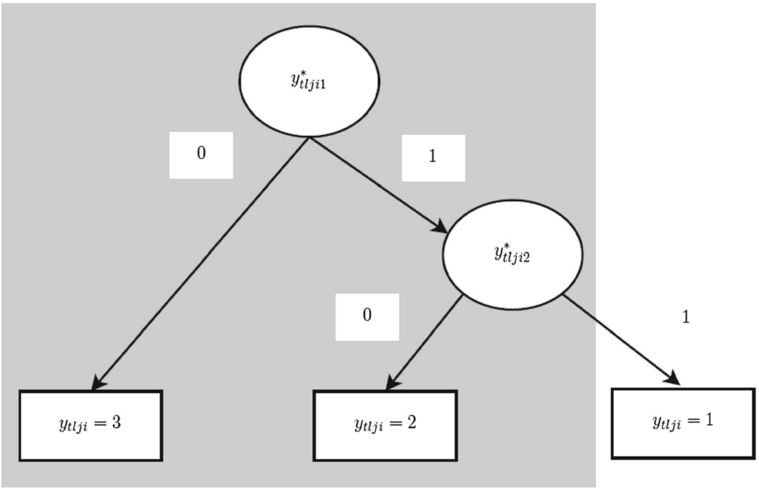

FIGURE 3.
Tree diagram for binary processes (two branches at each node in the tree) within a three-category paradigm; in the empirical study, Node 1 is for lexico-semantic processing and Node 2 is for ambiguity resolution; processes in the gray box are ignored in Cho et al. (2018).

TABLE 1.
An example of recoded responses: dynamic IRTree (top) and dynamic GLMM (bottom)

Response	Node r	y_{tlji}	y^*_{tljir}
Target	1	1	1
Competitor	1	2	1
Unrelated objects	1	3	0
Target	2	1	1
Competitor	2	2	0
Unrelated objects	2	3	NA

Response	y_{tlji}	u_{tlji}
Target	1	1
Competitor	2	0
Unrelated objects	3	0

NA indicates that $y_{tlji} = 3$ does not involve Node 2.

range = 18 for the number of time points). For example, as shown in Table 2 (bottom), there are observations for only 7 time points out of 10 at Node 2.

The regular approach of using the immediately preceding observation as a covariate works for Node 1 but not for Node 2 because of the necessarily missing observations. In this study, two different solutions are implemented. First, we use the last preceding observation which for Node 2 may not be the immediately preceding observation because of the missingness. This last preceding observation is still denoted with $t-1$ as a subscript although the literal $t-1$ observation may be missing. Second, lag covariates for the two nodes that refer directly to the three fixation options: target (T), competitor (C), and unrelated object (O) (see Table 2 for an illustration of the lag covariates). For the second approach, we use $x_{T(t-1)ljir}$ and $x_{C(t-1)ljir}$ to refer to the lag covariates instead of $y^*_{(t-1)ljir}$ to make them different from the first approach. The first lag covariate, $x_{T(t-1)ljir}$, indicates whether or not a person looked at the target at time point $t - 1$.

TABLE 2.

An illustration of covariate coding for autoregressive effects: time series over first ten time points for one trial of a given person and item for full data (top), data at Node 1 (middle), and data at Node 2 (bottom)

y_{tlji}	Time (ms)	$Time_{tljir}$	y^*_{tlji1}	y^*_{tlji2}
1(T)	180	0	1	1
2(C)	190	1	1	0
1(T)	200	2	1	1
3(O)	210	3	0	NA
3(O)	220	4	0	NA
2(C)	230	5	1	0
2(C)	240	6	1	0
3(O)	250	7	0	NA
1(T)	260	8	1	1
2(C)	270	9	1	0

y_{tlji}	Time (ms)	$Time_{tlji1}$	y^*_{tlji1}	x_{Ttlji1}	x_{Ctlji1}	$y^*_{(t-1)lji1}$	$x_{T(t-1)lji1}$	$x_{C(t-1)lji1}$
1(T)	180	0	1	1	0	–	–	–
2(C)	190	1	1	0	1	1	1	0
1(T)	200	2	1	1	0	1	0	1
3(O)	210	3	0	0	0	1	1	0
3(O)	220	4	0	0	0	0	0	0
2(C)	230	5	1	0	1	0	0	0
2(C)	240	6	1	0	1	1	0	1
3(O)	250	7	0	0	0	1	0	1
1(T)	260	8	1	1	0	0	0	0
2(C)	270	9	1	0	1	1	1	0

y_{tlji}	Time (ms)	$Time_{tlji2}$	y^*_{tlji2}	x_{Ttlji2}	x_{Ctlji2}	$y^*_{(t-1)lji2}$	$x_{T(t-1)lji2}$	$x_{C(t-1)lji2}$
1(T)	180	0	1	1	0	–	–	–
2(C)	190	1	0	0	1	1	1	0
1(T)	200	2	1	1	0	0	0	1
3(O)	210	3	NA	–	–	NA	–	–
3(O)	220	4	NA	–	–	NA	–	–
2(C)	230	5	0	0	1	1	0	0
2(C)	240	6	0	0	1	0	0	1
3(O)	250	5	NA	–	–	NA	–	–
1(T)	260	8	1	1	0	0	0	0
2(C)	270	9	0	0	1	1	1	0

1(T) indicates fixation to target (coded as 1); 2(C) indicates fixation to competitor (coded as 2); 3(O) indicates fixation to another object (coded as 3); – indicates a missing value; NA indicates that $y_{tljir} = 3$ does not involve Node 2; y^*_{tljir} is a recoded response; x_{Ttljir} is a covariate to indicate whether the observation was a target at a time point t for node r (coded as target = 1, otherwise = 0); x_{Ctljir} is a covariate to indicate whether the observation was a competitor at a time point t for node r (coded as competitor = 1, otherwise = 0).

The second lag covariate, $x_{C(t-1)ljir}$, indicates whether or not a person looked at the competitor at a time point $t - 1$. In this way, we have a meaningful lag covariate for all Node 2 observations even when the Node 2 observation at a time point $t - 1$ is missing. Although there are no missing observations for Node 1, the same lag covariates ($x_{T(t-1)ljir}$ and $x_{C(t-1)ljir}$) were considered for Node 1 so that both nodes can be treated in an equivalent way.

3.1.3. Model Specifications Below, we present two dynamic IRTree models: (a) the model with a lag covariate ($y^*_{(t-1)ljir}$) and (b) the model with two lag covariates ($x_{T(t-1)ljir}$ and $x_{C(t-1)ljir}$) to deal with missing observations at Node 2. The two models differ only with respect to how the AR parameters are defined. Having similar estimates and standard errors for the covariates of interest (i.e., experimental condition and trend effects) independent of the missing observation approach may indicate that they are not affected by the approach at Node 2. When estimates and standard errors for the covariates of interest did not differ between the two resulting models, we chose the first model (Eq. 1) as a simpler model.

A diagram of the dynamic IRTree model is first presented, and then, the equations are shown. A diagram of a dynamic IRTree model with a lag covariate $y^*_{(t-1)ljir}$ is presented in Fig. 2. Rectangles at Level 1 represent the (recoded) binary responses y^*_{tljir} for $t = 1, \ldots, 111$ (y_{0ljir} was dropped because we did not model it; see details in the subsection of initial time point treatment for $AR(1)$ effects below), and ovals represent random effects. The paths from the random effects ($\delta_{ljir}, \theta_{jr}, \beta_{jr}$) to y^*_{tljir}s ($t = 1, \ldots, 111$) indicate that dependency in y^*_{tljir} is explained by the random effects. The path from $time$ to the responses illustrates the fact that a fixed trend effect was considered in the model. The path from $y^*_{(t-1)ljir}$ to y^*_{tljir} indicates AR processes. The paths from the covariates $\mathbf{x}$ (except the trend and lag covariates, including experimental conditions, person covariates, and item covariates) to the random effects represent the fixed effects of the covariates $\mathbf{x}$, which shows that heterogeneity across trials, persons, and items is explained by the covariates $\mathbf{x}$. For a dynamic IRTree model with the two lag covariates ($x_{T(t-1)ljir}$ and $x_{C(t-1)ljir}$) a similar path diagram can be presented as in Fig. 2.

The model with a lag covariate $y^*_{(t-1)ljir}$. A dynamic IRTree model with a lag covariate $y^*_{(t-1)ljir}$ is written as follows:

$$\begin{aligned}\text{logit}[P(y^*_{tljir}|y^*_{(t-1)ljir}, time_{tljir}, \mathbf{x}, \delta_{ljir}, \lambda_{1jr}, \theta_{jr}, \lambda_{2ir}, \beta_{ir})] &= \eta_{tljir}\\ = [y^{*'}_{(t-1)ljir}\lambda_r + time^{'}_{tljir}\zeta_r + \mathbf{x}^{'}\boldsymbol{\gamma}_r] + \delta_{ljir} + [y^{*'}_{(t-1)ljir}\lambda_{1jr} + \theta_{jr}] &+ [y^{*'}_{(t-1)ljir}\lambda_{2ir} + \beta_{ir}],\end{aligned} \tag{1}$$

where t is an index for time ($t = 0, \ldots, T_{lji} - 1$) ($T_{lji} = T$ at Node 1), l is an index for trial ($l = 1, \ldots, L$), j is an index for person ($j = 1, \ldots, J$), i is an index for item ($i = 1, \ldots, I$), r is an index for node ($r = 1, \ldots, R$), $time_{tljir}$ is a linear time covariate for node r, and $\mathbf{x}$ is a design matrix of covariates (except the trend and lag covariates for node r) (i.e., the intercept, experimental conditions, person characteristics, item characteristics, and their interactions). Parameters in Eq. (1) are explained below:

- $\boldsymbol{\gamma}_r$ is a vector of node-specific fixed covariate effects.
- δ_{ljir}, θ_{jr}, and β_{ir} are design factor effects for node r. δ_{ljir} is a random trial effect to model dependency in responses due to the same trial number (Barr, 2008a), θ_{jr} is a random person effect (random person intercept) in order to allow for individual differences, and β_{ir} is a random item effect (random item intercept) to account for differences between the items. In psycholinguistics, simultaneously modeling person and item heterogeneity using crossed random person and item effects has been widely advocated (Baayen et al., 2008; Barr, 2008a; Jaeger, 2008; Quené & van den Bergh, 2008).
- ζ_r is a fixed trend effect for node r. This trend parameter reflects whether or not there is a trend over time in the fixation data. In principle, the trend effect may vary across trials and/or persons. In the model specification above, a *fixed* trend effect (i.e., a global [deterministic] linear trend) is considered to illustrate how the trend effect can be added to the dynamic IRTree model. We use a time variable as a covariate to model the trend effect as in multilevel modeling (e.g., Curran, Lee, Howard, Lane, & MacCallum, 2012).

- λ_r is a fixed $AR(1)$ effect for node r, λ_{1jr} is an $AR(1)$ random person effect (random person lag slope; person-specific deviation from λ_r) for person j and node r, and λ_{2ir} is an $AR(1)$ random item effect (random item lag slope; item-specific deviation from λ_r) for item i and node r. These AR parameters are modeled between adjacent binary responses as in the Markov chain of order 1 suggested by Cox (1970) for binary time series [also used in Bartolucci and Nigro (2010), Hung et al. (2008), and Jeon and Rabe-Hesketh (2016)] and in the AR latent trajectory model by Curran and Bollen (2001). The AR effect is allowed to be heterogeneous across persons and across items, because different persons may show different fixation strategies and items may also affect the fixation strategy. The $AR(1)$ effects can be interpreted as the log-odds ratio for the current response due to the previous response changing from 0 to 1.[2]

The terms within the first pair of square brackets represent the fixed effects of the model, the terms within the second pair of square brackets represent random person effects, and the terms within the third pair of square brackets represent random item effects. For $R = 2$ ($r = 1, 2$) as in our empirical study, random effects are assumed to have unstructured $D \times D$ variance–covariance matrices (where D is the dimension of the random effects), specified as follows. A random trial effect is used to model clustering by the same trial and is assumed to follow a multivariate normal distribution, $[\delta_{lji1}, \delta_{lji2}]' \sim MN(\mathbf{0}, \Sigma_{1(2\times 2)})$. Multivariate normality is assumed for random person effects and random item effects, respectively: $[\theta_{j1}, \theta_{j2}, \lambda_{1j1}, \lambda_{1j2}]' \sim MN(\mathbf{0}, \Sigma_{2(4\times 4)})$ and $[\beta_{i1}, \beta_{i2}, \lambda_{2i1}, \lambda_{2i2}]' \sim MN(\mathbf{0}, \Sigma_{3(4\times 4)})$. To identify the model, the means of all random effects are set to 0. Setting the means of the random effects to be 0 is sufficient when there are no scale parameters (e.g., item discrimination in item response models) as in the GLMM and design matrices for the fixed and random effects are full rank. In addition, it is necessary because random effects were included to model deviations from fixed effects. However, this is of course equivalent with the fixed effects being the means of the corresponding random effects.

The dynamic IRTree model specified in Eq. (1) can be reduced to a dynamic GLMM (e.g., Cho et al., 2018) when (a) a subscript r is dropped in parameters in Eq. (1) and (b) another kind of binary data coded as 1 for target fixations and 0 for non-target fixations (u_{tlji} as shown in Table 1 [bottom] and $u_{(t-1)lji}$) are used instead of y^*_{tljir} and $y^*_{(t-1)ljir}$ in Eq. (1).

The model with lag covariates, $x_{T(t-1)ljir}$ and $x_{C(t-1)ljir}$. Compared to Eq. (1), the model specification differs only for $AR(1)$ parameters. A dynamic IRTree model with the two lag covariates ($x_{T(t-1)ljir}$ and $x_{C(t-1)ljir}$) is written as:

$$
\begin{aligned}
&\text{logit}[P(y^*_{tljir}|x_{T(t-1)ljir}, x_{C(t-1)ljir}, time_{tljir}, \mathbf{x}, \delta_{ljir}, \lambda_{1jr}, \theta_{jr}, \lambda_{2ir}, \beta_{ir})] = \eta_{tljir}\\
&\quad = [x^{'}_{T(t-1)ljir}\lambda_{Tr} + x^{'}_{C(t-1)ljir}\lambda_{Cr} + time^{'}_{tljir}\zeta_r + \mathbf{x}^{'}\boldsymbol{\gamma}_r]\\
&\quad +\delta_{ljir} + [x^{'}_{T(t-1)ljir}\lambda_{T1jr} + x^{'}_{C(t-1)ljir}\lambda_{C1jr} + \theta_{jr}] + [x^{'}_{T(t-1)ljir}\lambda_{T2ir} + x^{'}_{C(t-1)ljir}\lambda_{C2ir} + \beta_{ir}],
\end{aligned}
\tag{2}
$$

where $\mathbf{x}$ is a design matrix of covariates (except the trend and lag covariates) (i.e., the intercept, experimental conditions, person characteristics, item characteristics, and their interactions), λ_{Tr} is an $AR(1)$ fixed effect for a target and node r, λ_{Cr} is an $AR(1)$ fixed effect for a competitor and node r, λ_{T1jr} is an $AR(1)$ person random effect (person random lag slope; person-specific deviation from λ_{Tr}) for a target, person j and node r, λ_{C1jr} is an $AR(1)$ person random effect (person random lag slope; person-specific deviation from λ_{Cr}) for a competitor, person j and node r, λ_{T2ir} is an $AR(1)$ item random effect (item random lag slope; item-specific deviation from λ_{Tr}) for a target, item i and node r, and λ_{C2ir} is an $AR(1)$ item random effect (item random lag slope;

[2] As explained below, we used deviation coding (-1 vs. 1) in model fitting, which compares each level to the grand mean. The interpretations of the $AR(1)$ effects are provided in the supplementary materials.

item-specific deviation from λ_{Cr}) for a competitor, item i and node r. The $AR(1)$ effects for the two lag covariates are for own-lag (e.g., T → T) and cross-lag (e.g., T → O) effects, which are presented in detail in Figure A.1 of the supplementary materials. A random trial effect to model clustering by the same trial numbers is assumed to follow a multivariate normal distribution, $[\delta_{lji1}, \delta_{lji2}]' \sim MN(\mathbf{0}, \Sigma_{1(2\times 2)})$. Multivariate normality is assumed for person random effects and item random effects for $R = 2$ as in our empirical study: $[\theta_{j1}, \theta_{j2}, \lambda_{T1j1}, \lambda_{T1j2}, \lambda_{C1j1}, \lambda_{C1j2}]' \sim MN(\mathbf{0}, \Sigma_{2(6\times 6)})$ and $[\beta_{i1}, \beta_{i2}, \lambda_{T2i1}, \lambda_{T2i2}, \lambda_{C2i1}, \lambda_{C2i2}]' \sim MN(\mathbf{0}, \Sigma_{3(6\times 6)})$. To identify the model, the means of all random effects are set to 0.

A simplification of the model is that the same random effects apply for the two nodes, so that the model is unidimensional in terms of nodes. This would imply that the processes do not depend on the node and therefore do not depend on the response option either. This simplification is not in line with the substantive theory at the basis of the study, but a comparison with the two-dimensional model is one way to test the multi-process hypothesis.

3.1.4. Coding for Trend and $AR(1)$ Effects For the trend effect, the group (by trial, person, and item)-mean centering of the time covariate was used within a node r. Lag covariates ($y^*_{(t-1)ljir}$, $x_{T(t-1)ljir}$, and $x_{C(t-1)ljir}$) were created for the unique combination of trial, person, and item. Deviation coding, −1 vs. 1, was used for the lag covariates (e.g., $y^*_{(t-1)ljir} = -1$ for 0; $y^*_{(t-1)ljir} = 1$ for 1).

3.1.5. Initial Time Point Treatment for $AR(1)$ Effects In AR modeling, the initial response variable (y^*_{0ljir}) is an issue because the corresponding covariate values do not exist. However, because there are often as many time points as 112 in the eye-tracking data, the effect of the starting point problem is minimal (Hsiao, 2003). It takes about 200 milliseconds to program and launch an eye movement in response to hearing the critical noun phrase, e.g., “small elephant” (Hallett, 1986; Salverda, Kleinschmidt, & Tanenhaus, 2014). Thus, we identify the onset of the critical adjective (e.g., “small”) on each individual trial and examine eye movements beginning 200 milliseconds after the onset of the adjective. In the present study, we use a 20 millisecond baseline (180–200 milliseconds) to facilitate handling a missing variable value y^*_{0ljir} at the 180 millisecond point; the observations at the first data point ($t = 0$) were omitted because there is no lag covariate value available.

3.2. Parameter Estimation

The marginal likelihood for Eq. (1) is written as:

$$\prod_{r=1}^{R}\prod_{j=1}^{J}\prod_{i=1}^{I}\int_{\boldsymbol{\zeta}_{1jr}}\int_{\boldsymbol{\zeta}_{2ir}}\left[\prod_{t=1}^{T_{lji}-1}\prod_{l=1}^{L}\left\{\int_{\delta_{ljir}} P(y^*_{tljir}|y^*_{(t-1)ljir}, time_{tljir}, \mathbf{x}, \delta_{ljir}, \lambda_{1jr}, \theta_{jr}, \lambda_{2ir}, \beta_{ir})g_1(\boldsymbol{\delta}_{ljir})d\boldsymbol{\delta}_{ljir}\right\}\right]$$
$$d\boldsymbol{\zeta}_{1jr}d\boldsymbol{\zeta}_{2ir}$$
$$\cdot\prod_{r=1}^{R}\prod_{j=1}^{J}\int_{\boldsymbol{\zeta}_{1jr}} g_2(\boldsymbol{\zeta}_{1jr})d\boldsymbol{\zeta}_{1jr}\cdot\prod_{r=1}^{R}\prod_{i=1}^{I}\int_{\boldsymbol{\zeta}_{2ir}} g_3(\boldsymbol{\zeta}_{2ir})d\boldsymbol{\zeta}_{2ir}, \quad (3)$$

where $\boldsymbol{\zeta}_{1jr} = [\theta_{j1}, \theta_{j2}, \lambda_{1j1}, \lambda_{1j2}]'$ for random person effects, $\boldsymbol{\zeta}_{2ir} = [\beta_{i1}, \beta_{i2}, \lambda_{2i1}, \lambda_{2i2}]'$ for random item effects, and $g_1(.)$, $g_2(.)$ and $g_3(.)$ are multivariate normal density functions. The random person effects and the random item effects are crossed random effects. The marginal likelihood involves integration over trials, persons, and items. The first term of the third line in Eq. (3) is the prior distribution of random person effects for marginalization, and the second term of the third line in Eq. (3) is the prior distribution of random item effects for marginalization.

In this study, parameter estimation was implemented using the glmer function in lme4 version 1.1.15 (Bates et al., 2018) R package (R Core Team, 2017). The glmer function for the logit link uses Laplace approximation corresponding to one adaptive quadrature point. It has been shown that Laplace approximation provides accurate estimates and standard errors in the GLMM with crossed random effects for intensive binary time-series data (e.g., Cho et al., 2018). However, Bates et al. (2018) noted in the lme4 manual that "warnings will occur even for apparently well-behaved fits with large data sets (p. 15)." Because the current study focuses on intensive (many time points) time-series data (e.g., 1,421,367 observations in our empirical study), convergence warnings were anticipated. Bates et al. (2018) suggest trying several optimizers in the glmer function as the gold standard when results are shown with convergence warnings. They consider the convergence warnings to be false positives when all optimizers converge to practically equivalent values. When convergence warnings are shown in this study, we check whether three optimizers, nloptwrap, bobyqa, and NelderMead in the glmer provide the same results. In addition, we check for singularities, which are defined as some of the constrained parameters of the random effects being on the boundary of the parameter space (i.e., equal to zero, or very close to zero).

3.3. *Model Selection, Testing, and Evaluation*

We first select a dynamic IRTree model among candidate models that have different random effect structures. We then test the experimental condition effects in the selected model. First, we investigate whether there are the same random effects for both nodes (unidimensional model) vs. different random effects depending on the node (multidimensional model). The unidimensional model implies that the same latent variable underlies the two nodes: fixations on the target or competitor vs. other and fixations on the target vs. the competitor. In testing whether there is just one dimension (versus two) across nodes, the trend and fixed $AR(1)$ effects for both nodes were included in the way as shown in the model specification below. The unidimensional models (Model A) are simplified versions of Eqs. (1) and (2), respectively, by dropping the r subscript in the parameters:

$$\eta_{tljir} = \gamma_1 + y^{*'}_{(t-1)ljir}\lambda + time^{'}_{tljir}\zeta + \delta_{lji} + \theta_j + \beta_i \tag{4}$$

and

$$\eta_{tljir} = \gamma_1 + x^{'}_{T(t-1)ljir}\lambda_T + x^{'}_{C(t-1)ljir}\lambda_C + time^{'}_{tljir}\zeta + \delta_{lji} + \theta_j + \beta_i, \tag{5}$$

where γ_1 is a fixed intercept parameter. Normality was assumed for the random effects in Eqs. (4) and (5). The multidimensional model (Model B) can be specified when node-specific parameters are considered in Eqs. (4) and (5):

$$\eta_{tljir} = \gamma_{1r} + y^{*'}_{(t-1)lji}\lambda_r + time^{'}_{tlji}\zeta_r + \delta_{ljir} + \theta_{jr} + \beta_{ir} \tag{6}$$

and

$$\eta_{tljir} = \gamma_{1r} + x^{'}_{T(t-1)ljir}\lambda_{Tr} + x^{'}_{C(t-1)ljir}\lambda_{Cr} + time^{'}_{tlji}\zeta_r + \delta_{ljir} + \theta_{jr} + \beta_{ir}. \tag{7}$$

Next, based on results of Model A vs. Model B, we explore whether random lag slopes are needed for persons and items to allow for heterogeneity in AR effects across persons and items, respectively.

For model selection, two information criteria, the marginal AIC (e.g., Greven & Kneib, 2010) and the Bayesian information criterion (BIC; Schwarz, 1978), were chosen. AIC is efficient in that it will asymptotically choose a model that minimizes the mean squared error of prediction, whereas BIC is consistent in that it will select the true model if the true model is among the candidate models (e.g., Burnham & Anderson, 2004). Thus, we consider both AIC and BIC to take into account efficiency and consistency. The candidate models were ranked from the smallest to the largest based on the AIC and BIC, and a best-fitting model was selected as the best combination of ranks.

For the selected model regarding random effects, experimental condition covariates (i.e., *Contrast* and *Privileged* covariates) were added to the model. Preliminary analyses show that there were convergence problems with random experimental condition effects across persons, which may indicate model over-fitting when the experimental condition effects do not vary much across persons. Thus, only fixed experimental condition effects were considered. Significance of the fixed effects at the 5% level using a two-tailed test was determined using a z-test.

For the final dynamic IRTree model (i.e., the selected model regarding random effects with the experimental condition covariates), residual analysis was conducted to check whether the final model provides an adequate description of the binary time series data (y^*_{tljir}). Model-based standardized residuals for time t, trial l, person j, and item i (i.e., $\frac{y^*_{tljir} - E(y^*_{tljir}|y^*_{(t-1)ljir}, time_{tljir}, \mathbf{x}, \delta_{ljir}, \lambda_{1jr}, \theta_{jr}, \lambda_{2ir}, \beta_{ir})}{\sqrt{Var(y^*_{tljir}|y^*_{(t-1)ljir}, time_{tljir}, \mathbf{x}, \delta_{ljir}, \lambda_{1jr}, \theta_{jr}, \lambda_{2ir}, \beta_{ir})}}$ for the dynamic IRTree model with a lag covariate $y^*_{(t-1)ljir}$, as an example) were calculated. Following standard practice for standardized residual analysis for large sample sizes (i.e., 111 time points, 288 trials, 152 subjects, and 96 items in our study), standard residuals smaller than -1.96 or larger than 1.96 were considered as indicating possible misfit at the 5% level. Further, Somers' rank correlation between the model-based probabilities ($P(y^*_{tljir}|y^*_{(t-1)ljir}, time_{tljir}, \mathbf{x}, \delta_{ljir}, \lambda_{1jr}, \theta_{jr}, \lambda_{2ir}, \beta_{ir})$ or $P(y^*_{tljir}|x_{T(t-1)ljir}, x_{C(t-1)ljir}, time_{tljir}, \mathbf{x}, \delta_{ljir}, \lambda_{1jr}, \theta_{jr}, \lambda_{2ir}, \beta_{ir})$) and the binary data y^*_{tljir} were calculated as a measure of the ordinal predictive power of the model.

4. Illustration

In this section, we illustrate the dynamic IRTree model using the eye-tracking data described earlier. The R code used in the application is shown in the supplementary materials.

4.1. Analysis Focus and Hypotheses

Our primary analysis focus is on detecting experimental condition effects. Interpreting trend effects is a secondary interest and was investigated in an exploratory way. A positive trend, while not of primary interest, is expected in the data because at the beginning of the time window, the listener does not know the identity of the target, but by the end of the time window, they have identified and fixated the target (Barr, 2008a; Brown-Schmidt 2009a; 2009b; 2012; Hanna et al., 2003). While a positive trend effect is expected for the use of the lexico-semantic information due to the rise in target fixations, there will always be many fixations to unrelated objects simply because there are so many of them. This will result in a smaller trend effect for the initial lexico-semantic processing that activates the target and competitor, compared to the ambiguity resolution process that distinguishes target from competitor. A strong AR effect is expected in the data because of the high sampling rate (100 hz in the downsampled data) and the fact that the eye moves comparatively slowly; thus, once the eye has fixated an object, it is likely to maintain fixation for many time points. The AR effect is considered as a controlling factor for the experimental condition effect in this study, although the AR effect itself can be of primary interest in other studies (e.g., Koval, Kuppens, Allen, & Sheeber, 2012; Kuppens, Allen, & Sheeber, 2010).

4.2. Characterizing Change Processes in Time-Series Data

In order to explore the trend and AR effects, logit-transformed proportion measures (called *empirical* logit) for *each* person j at each time point t ($ln\frac{P_{tjr}}{1-P_{tjr}}$ where $P_{tjr} = (\sum_l^L \sum_{i=1}^I y^*_{tljir})/LI$) and the empirical logit for *each* item i at each time point t ($ln\frac{P_{tir}}{1-P_{tir}}$ where $P_{tir} = (\sum_l^L \sum_{j=1}^J y^*_{tljir})/LJ$) were calculated based on binary responses y^*_{tljir} for each node in the tree. The top panel of Figure A.2 in the supplementary materials presents 112 box plots of the empirical logit across persons, over 112 time points (x-axis) at Node 1 (called a time-series plot). As shown in Figure A.2 (top), the overall trend was positive and linear over time. Fitted lines over time were similar between the linear function and Kernel-weighted local polynomial smoothing function, and only small deviations from the linear trend were observed (see Figure A.3 in the supplementary materials). To explore whether the trend pattern is similar across 288 trials graphically, we plot the empirical logit for *each* trial l at each time point t ($ln\frac{P_{tlr}}{1-P_{tlr}}$ where $P_{tlr} = \sum_{j=1}^J y^*_{tljir}/J$)[3] over 112 time points (x-axis) at each node. As shown in Figure A.4 in the supplementary materials, it was observed that the linear trend pattern is similar across the 288 trials in each node. The box plot of correlograms across persons presented in Figure A.2 (bottom) shows that the autocorrelations did not converge at zero, even at large values of lag. This pattern has also been observed when a time series also exhibits trend (Chatfield, 2004, p. 26). A similar pattern was found for items and at Node 2. Therefore, we include a linear trend effect in our illustration of the model. In addition, to investigate the order of AR, *partial* autocorrelations were calculated based on $ln\frac{P_{tjr}}{1-P_{tjr}}$ and $ln\frac{P_{tir}}{1-P_{tir}}$, respectively. The number of nonzero partial autocorrelations gives the order of the AR model (Chatfield, 2004). The partial correlations with the order of 1 are clearly larger than 0, and those with a larger lag are nearly 0. This suggests that only the $AR(1)$ needs to be included. Thus, the $AR(1)$ was considered in the dynamic IRTree model.

To investigate the linear trend over time (collapsing across trials) and $AR(1)$ effects simultaneously, linear growth $AR(1)$ models were fit based on $ln\frac{P_{tjr}}{1-P_{tjr}}$ for *each* person and $ln\frac{P_{tir}}{1-P_{tir}}$ for *each* item, respectively, using the `arima(1,0,0)` command in Stata (StataCorp, 2017). Inspection of the sample standard deviation (SD) of the estimates of the linear growth $AR(1)$ model indicates that heterogeneity occurred only for intercepts and $AR(1)$ estimates across persons and across items, respectively. In addition, according to sample SD of the linear growth $AR(1)$ model, little heterogeneity in trend estimates (SD of the trend estimates < 0.004) was observed across persons and items.

While separate analyses by trials, persons, and items at each node provide a description of the distinctive change processes, they ignore the full structure of the data across trials, persons, and items (shown in Fig. 2). Thus, we consider a dynamic IRTree model which accounts for change processes (linear trend and $AR(1)$) and heterogeneity across trials, persons, and items simultaneously in the dynamic IRTree model.

4.3. Model Selection and Model Fitting

Table 3 presents the log-likelihood (LL), AIC, and BIC for the 6 candidate models with $y^*_{(t-1)ljir}$ (special cases of Eq. 1) and 6 candidate models with $x_{T(t-1)ljir}$ and $x_{C(t-1)ljir}$ (special cases of Eq. 2) that were used to select the random slopes. First, the multidimensional models (Eqs. 6 and 7) having different random effects for each node (Model B in Table 3) can be compared

[3] Here, a subscript i is for a set of items which can be identified by a trial id and a person id.

TABLE 3.
Model selection of random effect models in dynamic IRTree and model fit for selected models: the model with $y^*_{(t-1)ljir}$ (A) and the model with $x_{T(t-1)ljir}$ and $x_{C(t-1)ljir}$ (B)

		Trial	Person		Item		Model selection		
Model	Node specific?	Trial int.	Person int.	Slope	Item int.	Slope	LL(NumP)	AIC	BIC
(A)									
Model A (Eq. 4)	No	✓	✓		✓		−92,311(6)	184,634[6]	184,707[6]
Model B (Eq. 6)	Yes	✓	✓		✓		−90,250(15)	180,531[5]	180,713[4]
Model B*	Yes	✓	✓		✓		−90,251(13)	180,527[4]	180,685[2]
Model B*-Person	Yes	✓	✓	✓	✓		−90,181(20)	180,403[2]	180,646[1]
Model B*-Item	Yes	✓	✓		✓	✓	−90,230(20)	180,500[3]	180,743[5]
Model B*-Person&Item	Yes	✓	✓	✓	✓	✓	−90,163(25)	180,380[1]	180,708[3]
(B)									
Model A (Eq. 5)	No		✓				−240,923(7)	481,859[6]	481,945[6]
Model B (Eq. 7)	Yes	✓	✓		✓		−87,892(17)	175,817[5]	176,024[3]
Model B*	Yes	✓	✓		✓		−87,892(15)	175,813[4]	175,996[1]
Model B*-Person	Yes	✓	✓	✓	✓		−87,813(33)	175,693[1]	176,094[2]
Model B*-Item	Yes	✓	✓		✓	✓	−87,871(33)	175,807[3]	176,209[4]
Model B*-Person&Item	Yes	✓	✓	✓	✓	✓	−87,802(51)	175,706[2]	176,326[5]

LL indicates a log-likelihood value; *NumP* indicates the number of parameters; *Int.* indicates an intercept; Model B* indicates that a trial random effect was considered at Node 2 only; numbers in square brackets indicate rank order of the AIC and BIC from the smallest to the largest; equations for Models B*s are presented in the supplementary materials; the penalty term regarding the number of parameters in the marginal AIC and BIC was calculated as the sum of the number of fixed parameters and the number of unique variance and covariance parameters for the random effects. In addition, the penalty term regarding sample size in the BIC was calculated as the total sample size, following the derivation shown in Cho and De Boeck (2018).

with the unidimensional models (Eqs. 4 and 5) having the same parameters between the two nodes (Model A in Table 3). AIC and BIC indicate that Model B fits better than Model A. In Model B, however, the variance estimate of the random trial intercept at Node 1 was near the boundary. Thus, the trial random intercept is allowed only at Node 2 (Model B* in Table 3). Next, random slopes for $AR(1)$ effects were added to Model B*. In Model B* with $y^*_{(t-1)ljir}$, AIC supported inclusion of random slopes for persons and items (Model B*-Person&Item in the top section of Table 3) as the best-fitting model. AIC supported inclusion of random slopes for persons as the second best-fitting model and BIC supported it as the best-fitting model (Model B*-Person in the top section of Table 3). Despite different suggestions by AIC and BIC, the Model B*-Person is the model which has the best combination of ranks (ranked second based on AIC and ranked first based on BIC). In Model B* with $x_{T(t-1)ljir}$ and $x_{C(t-1)ljir}$, AIC supported inclusion of random slopes for persons (Model B*-Person in the bottom section of Table 3). Although BIC did not support inclusion of random slopes, the Model B*-Person is also the model which has the best combination of ranks (ranked second based on AIC and ranked first based on BIC). Therefore, Model B*-Person was chosen and experimental condition covariates were added to Model B*-Person to test the effects of the conditions. Equations for Models B*s shown in Table 3 are presented in the supplementary materials. To confirm that pair clusters can be ignored in the dynamic IRTree model as in Cho et al. (2018), we fit the Model B*-Person with a random intercept for pairs. The variance of the random intercept was near 0 at each node, which indicates that dependence due to pairs can be ignored.

We use two condition contrasts with Helmert coding: (a) the *Contrast* covariate compares the One-Contrast condition (coded as -1) vs. the Two Contrasts-Privileged condition (coded as .5) and the Two Contrasts-Shared condition (coded as .5) and (b) the *Privileged* covariate directly compares the Two Contrasts-Privileged condition (coded as .5) and the Two Contrasts-Shared condition (coded as $-.5$) (the One-Contrast condition was coded as 0). In addition to the two experimental condition covariates (*Contrast* and *Privileged* covariates), person and item covariates were considered to explain heterogeneity among persons and items and to investigate their interaction effects with the lag covariate y^*_{tljir} or the two lag covariates $x_{T(t-1)ljir}$ and $x_{C(t-1)ljir}$. A mean-centered log-transformed *Cob* item covariate (frequency in spoken language) and a mean-centered *Ospan* person covariate (measure of working memory) were chosen, based on previous findings using the same data (Ryskin et al., 2015).

In testing the effects of experimental conditions, several models were considered. In Model 1, the *Contrast* and *Privileged* covariates were added to Model B*-Person with the two lag covariates $x_{T(t-1)ljir}$ and $x_{C(t-1)ljir}$. In Model 2, the *Contrast* and *Privileged* covariates were added to Model B*-Person with the lag covariate ($y^*_{(t-1)ljir}$). The estimates and standard errors for the covariates of interest (i.e., experimental condition and trend effects) did not differ between Model 1 and Model 2; thus, we chose Model 2 as a simpler model to test the focal effects. Further, the *Cob* item covariate and the *Ospan* person covariate were added to Model 1 and Model 2, respectively. Observations with missing covariate values (2.63% missingness for the *Ospan* person covariate and 1.16% missingness for the observations at the first time point) were discarded before fitting models using the `glmer` function. After comparison between Model 1 and Model 2 was made, we fit the model with the AR effect only or the trend effect only to show consequences of ignoring the AR effects or the trend effect on statistical inference for the experimental condition.

4.4. Results

The focal parameters were the experimental condition effects (the two contrasts), the trend effect, and the effects of person and item covariates. The *Cob* item covariate and the *Ospan* person covariate were not significant when they were added to Model 1 and Model 2, respectively. The log-likelihood value resulting from the three optimizers, `nloptwrap`, `bobyqa`, and `NelderMead` in `glmer`, was exactly the same. For Model 1 and Model 2, estimates and standard errors across the three optimizers were different in the second or third decimal point and standard errors were different in the third or fourth decimal point, indicating similar findings. In addition, there was no singularity problem in the three models and the estimated variance–covariance matrix ($\widehat{\Sigma}_2$ and $\widehat{\Sigma}_3$) was positive definite. Table 4 presents the estimates and the standard errors from Model 1 and Model 2 using the `nloptwrap` optimizer.

Standardized residuals for trial l, person j, and item i at time t were small. Somers' rank correlation was 0.985 for Model 1 and 0.984 for Model 2. Figure A.5 shows the model-data fit for a trial and a person by node, and a trial and an item by node for Model 2 for illustrative purposes. As shown in Figure A.5 (top), the average by-person proportion of binary responses y^*_{tljir} at each node (P_{tjr}) (i.e., data) is close to the average by-person predicted values at each node by Model 2 ($\bar{\eta}_{tjr} = (\sum_l \sum_i \tilde{\eta}_{tljir})/LI$) over time. In Figure A.5 (bottom), the average item proportion of binary responses y^*_{tlji} at each node (P_{tir}) is close to the average item predicted values at each node by Model 2 ($\bar{\eta}_{tir} = (\sum_l \sum_j \tilde{\eta}_{tljir})/LJ$) over time. A pattern similar to that shown in Figure A.5 was found for Model 1. Taken together, these findings indicate that Models 1 and 2 provide an adequate description of the data.

As shown in Table 4, statistical inference regarding the experimental condition and trend effects did not differ between Model 1 and Model 2. Below, we interpret the experimental condition, the trend, and the $AR(1)$ effects in Model 2.

4.4.1. Experimental Condition Effects The first condition contrast (*Contrast* covariate) compares the One-Contrast condition, in which the target can be immediately identified in the scene given the adjective, e.g., "small", compared to the two Two-Contrasts conditions, for which there is some ambiguity. A significant effect of the first contrast covariate at Node 1 (EST = 0.050, SE = 0.013, p value < 0.0002) is due to more looks to the Target and Competitor in the two Two-Contrasts conditions, compared to the One-Contrast condition. The ambiguity in the two Two-Contrasts conditions may have led to more fixations to those interest areas overall, compared to the One-Contrast condition in which only one of the two response options is consistent with the input. The first condition contrast was also significant at Node 2 (EST = −0.386, SE = 0.034, p value $< 2e-16$). This effect, now negative, was due to more target than competitor looks in the One-Contrast condition, compared to the two Two-Contrasts conditions. This effect is likely due to the fact that the context in the One-Contrast condition clearly identifies the Target, whereas there is more ambiguity between target and competitor in the two Two-Contrasts conditions.

The second condition contrast (*Privileged* covariate) directly compares the Two Contrasts-Shared condition with the Two Contrasts-Privileged condition and measures perspective-taking. If the listener uses information about the speaker's perspective, identification of the Target should be easier in the Two Contrasts-Privileged conditions. The perspective effect was not expected to differ from zero for Node 1 (Target and Competitor vs. Others) as the perspective effect should not be predictive at this node; indeed, it was very small and it was not significant (EST = 0.005, SE = 0.021, p value = 0.813). More surprising was the fact that the perspective effect was also not significant at Node 2 (Target vs. Competitor). The effect was in the predicted direction (EST = 0.071, SE = 0.046, p value = 0.123), due to more Target than Competitor looks in the Two Contrasts-Privileged condition (in which perspective can be used to identify the target) compared to the more ambiguous Two Contrasts-Shared condition. This is in contrast to other models of the same data which do find a significant perspective effect (Cho et al., 2018; Ryskin et al., 2015). A critical difference is that in those models, fixations to the Target vs. fixations to all other categories were modeled so that the nodes (processes) could not be differentiated. It may be that a target vs. else measure is more sensitive measure of perspective-taking than target vs. competitor in the dynamic IRTree model. This question deserves further inquiry.

TABLE 4.
Estimates (standard errors) of the Dynamic IRTree Model for an empirical study

	Model 1		Model 2	
	Node 1	Node 2	Node 1	Node 2
Fixed effects				
Intercept[γ_1]	**4.275**(0.023)	**1.086**(0.051)	**0.095**(0.021)	**1.313**(0.052)
$AR(1)$ylag[λ]	–		**4.181**(0.013)	**5.173**(0.031)
$AR(1)$T[λ_T]	**4.347**(0.014)	**2.751**(0.039)	–	
$AR(1)$C [λ_C]	**4.024**(0.016)	– **3.035**(0.030)	–	
Trend[ζ]	**0.005**(0.000)	**0.018**(0.001)	**0.006**(0.000)	**0.031**(0.001)
Privileged[γ_2]	– 0.006(0.021)	**0.033**(0.057)	0.005(0.021)	0.071(0.046)
Contrast[γ_3]	**0.074**(0.014)	– **0.338**(0.042)	**0.050**(0.013)	– **0.386**(0.034)

	Model 1						Model 2					
	SD	Corr					SD	Corr				
Random effects												
Trial (Σ_1)												
Node 1(δ_{lji1})	–						–					
Node 2(δ_{lji2})	0.001	–					0.094					
Person (Σ_2)												
Node 1: Intercept[θ_{j1}])	0.148						0.172					
Node 2: Intercept[θ_{j2}])	0.142	0.998					0.267	0.705				
Node 1: AR(1)ylag[λ_{1j1}]	–	–					0.115	– 0.360	– 0.180			
Node 2: AR(1)ylag[λ_{1j2}]	–	–	–				0.227	– 0.141	– 0.246	0.927		
Node 1: AR(1)T[λ_{T1j1}])	0.107	0.490	0.490	–			–	–	–	–		
Node 2: AR(1)T[λ_{T1j2}])	0.171	– 0.021	– 0.021	0.684	–		–	–	–	–	–	
Node 1: AR(1)C[λ_{C1j1}])	0.119	– 0.114	– 0.114	0.770	0.573		–	–	–	–	–	
Node 2: AR(1)C [λ_{C1j2}])	0.079	– 0.559	– 0.559	– 0.837	– 0.817	– 0.391	–	–	–	–	–	–
Item (Σ_3)												
Node 1: Intercept[β_{i1}])	0.123						0.127					
Node 2: Intercept[β_{i2}])	0.364	0.346					0.383	0.410				

– indicates that an effect is not modeled; values in bold indicate significance at the 5% level for fixed effects.

4.4.2. Trend Effects There was a significant (linear) trend effect per unit time at Node 1 (EST = 0.006, SE = 0.0002, p value $< 2e - 16$), reflecting increased looks to the target and competitor (e.g., small elephant, small envelope) as the linguistic expression unfolds. The trend estimate of 0.006 means that per unit of time there is 0.0015 increase in probability with 0.50 as the reference probability ($= \frac{1}{1+exp(-[0.006])} - \frac{1}{1+exp(0)}$) or 1.002 ($= exp(0.0015)$) increase in odds ratio for target and competitor vs. other fixations (controlling for the other covariates). There was also a significant trend effect per unit time at Node 2 (EST = 0.031, SE = 0.0008, p value $< 2e - 16$); this larger magnitude effect reflects the fact that the preference to fixate the target (e.g., small elephant) more than the competitor (e.g., small envelope) increases more over time. The trend estimate of 0.031 indicates that per unit of time there is 0.008 increase in probability with 0.50 as the reference probability ($= \frac{1}{1+exp(-[0.031])} - \frac{1}{1+exp(0)}$) or 1.008 ($= exp(0.008)$) in odds ratio for target vs. competitor fixations (controlling for the other covariates).[4]

4.4.3. AR(1) Effects The fixed $AR(1)$ effects (as a controlling factor) were significant at Node 1 (EST = 4.181, SE = 0.013, p value $< 2e - 16$) and at Node 2 (EST = 5.173, SE = 0.031, p value $< 2e - 16$). With the deviation coding for a lag covariate (i.e., $y^*_{(t-1)ljir} = -1$ vs. 1), the AR estimates of 4.181 and 5.173 are 2.091 ($= 4.181/2$; $exp(2.091) = 8.089$ odds ratio) and 2.587 ($= 5.173/2$; $exp(2.587) = 13.283$ odds ratio), respectively, controlling for the other covariates. These large effect sizes of the AR effects indicate that there are strong carryover effects: (a) Target or Competitor at time point $t - 1 \rightarrow$ Target or Competitor at time point t and (b) Others at time point $t - 1 \rightarrow$ Others at time point t at Node 1; (a) Target at time point $t - 1 \rightarrow$ Target at time point t and (b) Competitor at time point $t - 1 \rightarrow$ Competitor at time point t at Node 2. There was non-ignorable variability in the $AR(1)$ effects across persons ($SD = 0.115$).

5. Simulation Study

5.1. Simulation Design and Analysis

A simulation study was designed to answer the following questions when the same conditions of the empirical study were considered (i.e., 112 time points, 288 trials, 152 persons, and 96 items): (a) can the parameters of the selected model (Model 2 in Table 4) be recovered at a satisfactory level?; (b) what are the consequences of ignoring trend and AR, AR or trend on the experimental condition effects in the presence of change processes?; (c) what are the consequences of modeling change processes (trend and AR parameters) for the experimental condition effects when they do not exist; (d) what are the consequences for ignoring a complex trend (quadratic trend) on the experimental condition effects and linear trend effects in the presence of the complex trend? Questions (b)–(d) investigate model misspecifications regarding the change processes. The same true model (Model 2 in Table 4) was considered for questions (a) and (b), and a modified Model 2 was considered for questions (c) and (d), respectively.

To answer questions (a) and (b), estimates of Model 2 (reported in Table 4) were considered true parameters. For question (a), Model 2 was fitted to the simulated data set. For question (b), Model 2 without a fixed trend effect and without $AR(1)$ fixed and random effects (called Model 2-1), Model 2 without $AR(1)$ fixed and random effects (called Model 2-2), and Model 2 without a fixed trend effect (called Model 2-3) were fitted to the same simulated data sets generated based on Model 2. Regarding question (c), estimates of Model 2 (reported in Table 4) without a fixed trend effect and without $AR(1)$ fixed and random effects (called Model 3) are

[4] Across the entire 111 time points (controlling for the other covariates), the trend estimate of 0.006 at Node 1 means that there is 1.174 ($= exp(0.161)$) increase in odds ratio for the target and competitor fixation and the trend estimate of 0.031 indicates that there is 1.598 ($= exp(0.469)$) in odds ratio for the target fixation at Node 2.

considered the true parameters. To investigate the consequences of modeling change processes when they do not exist, Model 2 was fitted to the simulated data. To answer question (d), Model 2 with an additional quadratic time effect was fitted to the empirical data of the current study (Model 4), and estimates of Model 4 (reported in Table A.1 of the supplementary materials) were considered the true parameters. As shown in Table A.2, there were significant quadratic trend effects although the effects were small. To show the consequences of ignoring the quadratic time effect on experimental conditions and the linear trend, Model 2 was fitted to the simulated data set under Model 4.

Laplace approximation implemented in the `glmer` function is used for parameter estimation. The `nloptwrap` optimizer in the `glmer` function was chosen, as used in the empirical study. For each research question, two hundred replications for model fitting were considered. For all questions (a)–(d), bias and root-mean-square error (RMSE) of estimates were calculated for parameter recovery and the mean standard error estimates (M(SE)) across 200 replications were compared with the standard deviations (SD) of the estimates for evaluation of standard error accuracy. There were no convergence problems for any of the replications, although there were the same warning messages observed in the empirical data set.

5.2. *Results*

Table 5 presents results for the questions (a) and (b). Bias for trend and experimental condition effects in Model 2 was very close to 0. RMSE for Model 2 was comparable to those observed in GLMM for binary responses in the case of a large number of persons and items (e.g., Cho et al., 2018). RMSE tends to be larger at Node 2 than at Node 1, which may be due to the fact that the number of observations is smaller at Node 2 than at Node 1. In addition, for fixed effects in Model 2, the M(SE) across 200 replications approached the SD of the estimates, indicating that the estimated standard errors are approximately correct. When trend and AR effects were ignored as in Model 2-1 (misspecified model) and AR effects were ignored as in Model 2-2 (misspecified model), bias and RMSE of Model 2-1 and Model 2-2 (misspecified models) were larger than those of Model 2 (true model) and standard errors for the two experimental condition effects were underestimated. The underestimated standard errors lead to an inflated Type I error rate and to an overestimated power rate. Our simulation study results indicate that the inferential biases for the fixed effects of interest are reduced by modeling change processes. When the fixed trend effect is ignored in Model 2-3 (misspecified model), bias and RMSE of the two experimental condition effects at Node 1 were the same as those at Node 1 under Model 2 (true model). However, the estimates for the two experimental condition effects at Node 2 were more biased under Model 2-3 compared to Model 2, although a noticeable underestimation was not found for the standard errors of the effects. As presented in Table A.3 for question (c), bias, RMSE, SD, and M(SE) of Model 3 (true model) and Model 2 (misspecified model) were comparable. This result suggests that modeling change processes (trend and AR) did not distort results for the experimental condition effects when there were no change processes. Based on the results reported in Table A.4 for question (d), bias, RMSE, SD, and M(SE) were similar between Model 4 (true model) and Model 2 (misspecified model) for the experimental condition and linear trend effects, although a larger bias and RMSE for the intercept and AR fixed effects were found in Model 2. This result implies that ignoring the quadratic trend effect (small deviations from the linear function as in our empirical study) did not affect results for the experimental condition and the linear trend effects.

To sum up, the results of the simulation study for our empirical illustration demonstrate that the model specified in the current study was found to accurately recover the parameters when the specified model is the data generating model. Furthermore, the simulation study shows that in order to accurately estimate the experimental condition effects, it is necessary to model change

TABLE 5.
Results of the dynamic IRTree model for the simulation study: questions (a) and (b)

	Model 2 True				Model 2-1 Misspecified			
	Bias	RMSE	SD	M(SE)	Bias	RMSE	SD	M(SE)
Fixed effects								
Node 1: Intercept[γ_{11}]	0.001	0.021	0.021	0.021	-0.367	0.368	0.033	0.030
Node 2: Intercept[γ_{12}]	0.002	0.054	0.054	0.051	-1.210	1.211	0.045	0.054
Node 1: $AR(1)$ylag[λ_1]	-0.001	0.012	0.012	0.013	–			
Node 2: $AR(1)$ylag[λ_2]	-0.005	0.034	0.034	0.033	–			
Node 1: Trend[ζ_1]	0.000	0.000	0.000	0.000	–			
Node 2: Trend[ζ_2]	0.000	0.001	0.001	0.001	–			
Node 1: Privileged[γ_{21}]	0.004	0.019	0.019	0.020	0.005	0.027	0.027	0.005
Node 2: Privileged[γ_{22}]	0.000	0.037	0.037	0.035	-0.021	0.047	0.042	0.008
Node 1: Contrast[γ_{31}]	0.000	0.012	0.012	0.013	0.010	0.020	0.017	0.003
Node 2: Contrast[γ_{32}]	0.001	0.020	0.020	0.022	0.135	0.136	0.023	0.005
Random effects								
Trial (Σ_1)								
Node 2: Intercept[δ_{lji2}]	0.000	0.006			0.081	0.082		
Person (Σ_2)								
Node 1: Intercept[θ_{j1}])	0.000	0.004			0.042	0.043		
Node 2: Intercept[θ_{j2}])	-0.002	0.020			0.110	0.112		
Node 1: AR(1)ylag[λ_{1j1}]	0.000	0.003			–			
Node 2: AR(1)ylag[λ_{1j2}]	0.008	0.020			–			
Covariances	0.000	0.006			0.042	0.044		
Item (Σ_3)								
Node 1: Intercept[β_{i1}])	0.001	0.003			0.022	0.022		
Node 2: Intercept[β_{i2}])	0.001	0.023			-0.009	0.020		
Covariance	0.002	0.006			0.017	0.018		

TABLE 5.
continued

	Model 2-2 Misspecified				Model 2-3 Misspecified			
	Bias	RMSE	SD	M(SE)	Bias	RMSE	SD	M(SE)
Node 1: Intercept[γ_{11}]	−0.368	0.369	0.033	0.030	−0.003	0.022	0.022	0.021
Node 2: Intercept[γ_{12}]	−1.207	1.207	0.046	0.056	−0.081	0.094	0.049	0.044
Node 1: AR(1)ylag[λ_1]	–				−0.030	0.032	0.012	0.012
Node 2: AR(1)ylag[λ_2]	–				−0.196	0.199	0.032	0.029
Node 1: Trend[ζ_1]	−0.009	0.009	0.000	0.000	–			
Node 2: Trend[ζ_2]	−0.016	0.016	0.001	0.000	–			
Node 1: Privileged[γ_{21}]	0.005	0.027	0.027	0.005	0.004	0.019	0.019	0.020
Node 2: Privileged[γ_{22}]	−0.019	0.047	0.043	0.008	−0.006	0.031	0.031	0.035
Node 1: Contrast[γ_{31}]	0.010	0.020	0.017	0.003	0.000	0.012	0.012	0.013
Node 2: Contrast[γ_{32}]	0.126	0.129	0.024	0.005	0.080	0.081	0.019	0.022
Random effects								
Trial (Σ_1) Node 2: Intercept[δ_{lji2}]	0.087	0.087			−0.009	0.009		
Person (Σ_2)								
Node 1: Intercept[θ_{j1}])	0.042	0.043			0.001	0.005		
Node 2: Intercept[θ_{j2}])	0.122	0.124			−0.017	0.024		
Node 1: AR(1)ylag[λ_{1j1}]	–				−0.002	0.003		
Node 2: AR(1)ylag[λ_{1j2}]	–				−0.010	0.016		
Covariances					0.002	0.008		
Item (Σ_3)								
Node 1: Intercept[β_{i1}])	0.022	0.023			0.001	0.003		
Node 2: Intercept[β_{i2}])	0.000	0.019			−0.050	0.052		
Covariance	0.018	0.020			0.000	0.005		

Model 2-1 is a misspecified model in which the AR and trend effects were ignored; Model 2-2 is a misspecified model in which the AR effects were ignored; Model 2-3 is a misspecified model in which the trend effect was ignored; – indicates that an effect is not modeled; *SD* indicates the standard deviations of the estimates across 200 replications; *SE* indicates the mean standard error estimates across 200 replications, which are available for the fixed effects; average bias and RMSE across covariances of random person effects are reported.

processes (trend and AR parameters or AR parameters) in the data when there is a small trend effect and a large $AR(1)$ effects as in our empirical study. However, modeling change processes did not affect results for experimental condition effects when the processes did not exist in the data. Lastly, ignoring a small quadratic trend effect did not affect the interpretations of results for experimental condition and linear trend effects.

6. Summary, Discussion, and Conclusion

6.1. Summary

This study presents a dynamic IRTree model which can be used to model change processes (trend and AR) and which facilitates the decomposition of data into various sources of heterogeneity (trial, person, and item effects). We illustrate how this model can be used to advance the analysis of eye-tracking time-series data through the exploration of the cognitive processes that guide selection among competing response options. By applying the dynamic IRTree model to visual-world eye-tracking data, the researcher can now ask and answer central questions about how multinomial processes guide real-time language understanding in a way that was not technically feasible before. These questions include issues related to how distinct cognitive processes lead to different consideration of competing response options, and further, how these processes take into account different sources of information, including disambiguation and semantic information. In the present study, we model fixations to language-relevant (vs. irrelevant) items at Node 1, and we model fixations to the target vs. competitor at Node 2. By focusing on different processes at each node, this allows us to separately ask questions about covariates that are assumed to affect linguistic processing in general (Node 1 effects), vs. questions about covariates that are assumed to affect the ability to select the target over the competitor specifically (Node 2 effects). This dramatically increases the precision of hypotheses that can be tested. It undoes the confounding due to coding the responses just in terms of target fixations vs. non-target fixations.

In the illustrative data set, change processes were first explored with time-series plots, descriptive statistics (i.e., autocorrelation and partial autocorrelation) and the linear growth $AR(1)$ model based on the empirical logit; then, the patterns of change processes from the exploratory analyses were modeled in the dynamic IRTree model. We found that the model with trend and $AR(1)$ adequately described the data. In this model, the trend effect is small and the AR effects are large. Due to the small trend effect, statistical inference on experimental condition effects (which is of primary interest) did not change when the trend effects are ignored. While the trend effect is not the empirical focus, a trend is expected given that the participant identifies the target over the course of the time window.

In the simulation study, we show that misspecification of the trend and the $AR(1)$ led to biased estimates for the effects of interest (i.e., fixed experimental condition effects in our application). Given the simulation conditions from the empirical study in which there was a small trend effect and large $AR(1)$ effects, ignoring the $AR(1)$ resulted in more biased estimates and underestimated standard errors than ignoring the trend effect. Given these findings, we recommend researchers exploring change processes using the time-series plots and descriptive statistics (i.e., autocorrelation and partial autocorrelation) (as illustrated in the empirical study), prior to applying the dynamic IRTree model to other eye-tracking data sets. A guideline for exploring the change processes is presented in Figure A.6 of the supplementary materials. In addition, after a dynamic IRTree model is selected for the change processes, model evaluation is an important step to check whether the model describes the data adequately, as we showed in the current study using residual analysis and Somers' rank correlation.

6.2. Discussion

What did we learn from node differentiation using the dynamic IRTree model?

A crucial aspect of the tree approach is that it makes a distinction between processes occurring at two nodes: one node for the processing of lexico-semantic information and another for distinguishing the target referent from the competitor using cues such as ambiguity resolution. These two processes are confounded without using a tree approach. Disentangling the processes occurring at the two nodes allows for differentiation of these processes, affording novel and more theoretically specific hypothesis tests compared to a non-tree approach. We illustrate this in the following:

- The nodes correspond to two different dimensions (two different types of language processing). One refers to the lexico-semantic process of interpreting the initial words in the phrase, e.g., "the small e-...", the other refers to the resolution of ambiguity between the target and competitor. Model fit results and correlations between the two random person intercepts confirm that each node (each process) comes with its own dimension.[5] The hypothesized distinction between initial lexical activation processes and those processes which resolve competition among the lexical candidates is a property of several models of language processing (e.g., Barr, 2008b; McMurray et al., 2010; 2019); here, we illustrate a method for distinguishing between those processes in a single model.
- At Node 1 we observe that the *Contrast* effect was negative (−0.386), indicating that the lexico-semantic processing of the unfolding expression, e.g., the "small e-", activates the target and competitor more in the two *Two-Contrasts* conditions than in the *One-Contrast* condition. This results from the fact that while the target is always consistent with the unfolding expression, the competitor object is consistent with it only in the two *Two-Contrasts* conditions. The contrast effect was positive (0.050) at Node 2, which zeroed in on the competition between target and competitor. Because the *One-Contrast* condition most clearly picks out the target as the intended referent, there were more target looks in the *One-Contrast* condition compared to the two *Two-Contrasts* conditions. Thus, depending on node, the direction of the Contrast effect flipped. This finding raises the novel possibility that contexts with relatively little ambiguity (e.g., *One-Contrast* condition) result in overall less referential activation compared to contexts with more ambiguity (e.g., *Two-Contrasts* conditions), despite facilitating target identification. If so, this observed trade-off between referential activation and target identification makes novel predictions about how other factors, such as sentence predictability and language production constraints (Altmann & Kamide, 1999; MacDonald, 2013), would differentially affect lexical activation and ambiguity resolution processes. This result reflects the kind of novel finding that would not be possible without a tree-type approach.
- The trend effect differs depending on the node. The trend is clearly steeper for the ambiguity resolution process than for lexico-semantic processing (0.031 vs. 0.006). The steeper trend in ambiguity resolution likely relates to the fact that by the end of the time window, there is a strong preference to fixate the target over the competitor. By contrast, a smaller trend effect in the lexico-semantic processing is likely due to the fact that the preference to fixate the target and competitor over the unrelated objects is weakened by the fact that there are simply many unrelated objects on screen, and together they attract many fixations. Isolating the trend effect by node allows the researcher to investigate novel questions about how the distinct processes by node result in distinct activation profiles for the candidate referents.

[5] When Model 2 from Table 4 is estimated with just one random person intercept for both nodes, model fit is poorer (AIC = 181,432, BIC = 181,627), compared to the full Model 2 in Table 2 which has two random person intercepts, one per node (AIC = 180,262, BIC = 180,554). The correlation between the two random person intercepts is even smaller for Model 1 than for Model 2 (.414 vs. .705).

These three results show clearly that it is informative and important to undo the confounding inherent to approaches which do not make use of a tree approach. The tree approach of decomposing response options into nodes is a general approach that can be used for all cases with more than two response options (see De Boeck & Cho, 2019, for examples).

Methodological limitations and discussions

In the current study, we focus on model specification and illustration using an empirical data set. Thus, the following methodological limitations remain. First, a dynamic IRTree model was presented for intensive (many time points) time-series data (1,421,367 observations in our empirical study) with various random effects; as a result, parameter estimation involves high-dimensional integration. For these reasons, the amount of computing required was large. For Model 2 in our empirical study, about 7 hours (user time in R) were required on a 2.81GHz computer with 16.0 GB of RAM. Substantial computing resources or development of speedier algorithms is required to improve the feasibility of implementation.

Second, it has been observed that results from the `glmer` function are shown with convergence warning messages (not error messages) for large data sets (Bates et al., 2018, p. 15). To demonstrate that parameter estimates and standard errors are reliable in the empirical study, we checked that the three optimizers, `nloptwrap`, `bobyqa`, and `NelderMead` in the `glmer`, provided the same results. Furthermore, to assure that results from the `glmer` function are reliable, we checked that estimates and statistical inference for fixed effects of interest are comparable between Laplace approximation and Bayesian analysis implemented in `Stan` (Carpenter et al., 2017). The detailed implementation of Bayesian analysis and comparability between the two estimation methods (see Table A.5) were shown in the supplementary materials. Although we show that results reported in the current case study are reliable, future work is needed to generalize our findings to other data structures differing in the number of data modes (time, trial, person, and item) and the magnitude of effects.

Third, selection of random effects (random slopes) is an important step for valid statistical inferences on fixed effects in GLMM (e.g., Barr, Levy, Scheepers, & Tily, 2013). In this study, marginal AIC and BIC were used to account for efficiency and consistency in the use of (approximate) marginal maximum likelihood estimation. For the random slope models, AIC and BIC did not provide perfectly consistent results. We chose the best-fitting model as the model having the best combination of AIC and BIC. For the final model, we used to investigate the effects of experimental conditions (the model with $y^*_{(t-1)ljir}$), BIC was the smallest and AIC was the second smallest among candidate models having different random effect structures. Bringmann et al. (2017) found that BIC outperformed the AIC for generalized additive models when there are over 100 time points. However, it is unclear whether this finding can be generalized to GLMMs.

Fourth, we expected significant experimental condition contrast effects in the empirical study. In an analysis of the same data with a binary time-series model for target vs. non-target fixations, the two condition contrasts (One-Contrast vs. Two-Contrasts, Two-Contrasts-Privileged vs. Two-Contrasts-Shared) both had a significant effect on target fixations. Using the dynamic IRTree model with two nodes (target or competitor vs. unrelated object, target vs. competitor), only the first of the two contrasts (One-Contrast vs. Two-Contrasts) had a significant effect, in favor of target and competitor vs. other (Node 1) and in favor of the target vs competitor (Node 2). However, the second condition contrast (Two-Contrasts-Privileged vs. Two-Contrasts-Shared) was neither significant at Node 1 nor at Node 2. It is important to note that the standard error of the estimated effect at Node 2 in the dynamic IRTree model (0.046) is much larger than for the corresponding effect in the model of binary target fixations (0.010). A possible explanation of this discrepancy is due to the correction for underestimated errors by accounting for a large number of random effects in the dynamic IRTree approach.

Fifth, two kinds of errors have been discussed in using lag covariates ($y^*_{(t-1)ljir}$, $x_{T(t-1)ljir}$, and $x_{C(t-1)ljir}$) (Hamaker & Grasman, 2015). The first kind of error is the error in the sample mean estimates of the true mean when centering is considered for AR effects, called Nickell's bias (Nickell, 1981). The bias for the AR effect is not expected in the uncentered model we used. The second kind of error is measurement error in the covariates. It has been found that the covariate effects and their standard errors can be biased in the presence of measurement error in covariates (e.g., Lüdtke et al., 2008). However, the lag covariates refer to previous responses; they are not a measurement of external variables with a possibly imperfect reliability. It is an inherent AR modeling feature of modeling direct dependencies between observed responses that the observed responses themselves are modeled as a function of earlier observed responses.

6.3. Conclusion

Intensive polytomous time-series data can be fruitfully analyzed by taking into consideration the multiple cognitive processes that are involved. Here we consider the case of eye-tracking data, which provide a dense and rich source of information about multiple real-time cognitive processes in a wide variety of content domains. Time-related analysis of eye gaze as in our empirical study was popularized by the introduction of the visual-world eye-tracking technique, in which a participant produces or interprets spoken or signed language while viewing an associated visual scene. In the case of spoken language comprehension, the speech is an external stimulus that drives eye fixations to language-relevant interest areas in the scene (Tanenhaus et al., 1995). The visual-world paradigm has emerged as a dominant technique for the study of language processing, the original article has been cited over 2500 times, and this elicited-gaze technique has been extended to multiple literatures including philosophy (Sedivy, 2007), child developmental disorders (McMurray et al., 2010), cross-cultural psychology (Barrett et al., 2013), human–computer interaction (Qu & Chai, 2007), human–robot interaction (Staudte & Crocker, 2011), and surgery and medical education (Merali, Veeramootoo, & Singh, 2019). Despite the ubiquity and utility of eye-gaze time-series data as a measure of cognitive processes, current data analytic practices ignore multinomial processing. In spite of methodological limitations, to the best of our knowledge, this paper is the first attempt in the literature to advance knowledge of multiple ongoing cognitive processes in tasks where eye gaze provides a window into cognition. As we have shown, it is important to differentiate between the distinct cognitive processes that drive consideration among the competing response options. Making use of a response tree is useful for such a differentiation. The new method that we have introduced will allow researchers to differentiate hypothesized cognitive processes and test distinct predictions regarding the cognitive mechanisms underlying each process. In sum, this new methodology for analyzing eye-tracking data will allow researchers—across a variety of fields using variants of the visual-world paradigm—to make better use of the rich source of information that the eyes provide about cognition.

Acknowledgments

We thank an associate editor and five anonymous reviewers for their helpful comments.

Funding Funding was provided in part by the National Science Foundation (SES 1851690) to Sun-Joo Cho, Sarah Brown-Schmidt, and Paul De Boeck. The original data collection was supported in part by the National Science Foundation (BCS 1257029 and BCS 1556700) to Sarah Brown-Schmidt. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*, 419–439. https://doi.org/10.1006/jmla.1997.2558.

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*, 247–264. https://doi.org/10.1016/S0010-0277(99)00059-1.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412. https://doi.org/10.1016/j.jml.2007.12.005.

Barr, D. J. (2008a). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, *59*, 457–474. https://doi.org/10.1016/j.jml.2007.09.002.

Barr, D. J. (2008b). Pragmatic expectations and linguistic evidence: Listeners anticipate but do not integrate common ground. *Cognition*, *109*, 18–40. https://doi.org/10.1016/j.cognition.2008.07.005.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278. https://doi.org/10.1016/j.jml.2012.11.001.

Barrett, H. C., Broesch, T., Scott, R. M., He, Z., Baillargeon, R., & Wu, D. (2013). Early false-belief understanding in traditional non-western societies. *Proceedings of the Royal Society of London B: Biological Sciences*, *280*, 2012–2654. https://doi.org/10.1098/rspb.2012.2654.

Bartolucci, F., & Nigro, V. (2010). A dynamic model for binary panel data with unobserved heterogeneity admitting a n-consistent conditional estimator. *Econometrica*, *78*, 719–733. https://doi.org/10.3982/ecta7531.

Batchelder, W. H., & Crowther, C. S. (1997). Multinomial processing tree models of factorial categorization. *Journal of Mathematical Psychology*, *41*, 45–55. https://doi.org/10.1006/jmps.1997.1146.

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, *6*, 57–86. https://doi.org/10.3758/bf03210812.

Bates, D., Mächler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., et al. (2018). *Package "lme4": Linear mixed-effects models using 'eigen' and s4*. Retrieved March 10, 2018 from https://cran.r-project.org/web/packages/lme4/lme4.pdf.

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods*, *17*(4), 665–678. https://doi.org/10.1037/a0028111.

Böckenholt, U. (2017). Measuring response styles in Likert items. *Psychological Methods*, *22*, 69–83. https://doi.org/10.1037/met0000106.

Bringmann, L. F., Hamaker, E. L., Vigo, D. E., Aubert, A., Borsboom, D., & Tuerlinckx, F. (2017). Changing dynamics: Time-varying autoregressive models using generalized additive modeling. *Psychological Methods*, *22*, 409–425. https://doi.org/10.1037/met0000085.

Brown-Schmidt, S. (2009a). The role of executive function in perspective-taking during on-line language comprehension. *Psychonomic Bulletin & Review*, *16*, 893–900. https://doi.org/10.3758/pbr.16.5.893.

Brown-Schmidt, S. (2009b). Partner-specific interpretation of maintained referential precedents during interactive dialog. *Journal of Memory and Language*, *61*, 171–190. https://doi.org/10.1016/j.jml.2009.04.003.

Brown-Schmidt, S. (2012). Beyond common and privileged: Gradient representations of common ground in real-time language use. *Language and Cognitive Processes*, *27*, 62–89. https://doi.org/10.1080/01690965.2010.543363.

Brown-Schmidt, S., Gunlogson, C., & Tanenhaus, M. K. (2008). Addressees distinguish shared from private information when interpreting questions during interactive conversation. *Cognition*, *107*, 1122–1134. https://doi.org/10.1016/j.cognition.2007.11.005.

Browne, M. W., & Nesselroade, J. R. (2005). Representing psychological processes with dynamic factor models: Some promising uses and extensions of autoregressive moving average time series models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Multivariate applications book series. Contemporary psychometrics: A festschrift for Roderick P. McDonald* (pp. 415–452). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Burnham, K. P., & Anderson, D. R. (2004). *Model selection and multimodel inference: A practical information-theoretical approach* (2nd ed.). New York, NY: Springer. https://doi.org/10.1007/b97636.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., et al. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*, 1–32. https://doi.org/10.18637/jss.v076.i01.

Chatfield, C. (2004). *The analysis of time series: An introduction* (6th ed.). London: Chapman Hall. https://doi.org/10.1007/978-1-4899-2921-1.

Cho, S.-J., & De Boeck, P. (2018). A note on N in Bayesian information criterion for item response models. *Applied Psychological Measurement*, *42*, 169–172. https://doi.org/10.1177/0146621617726791.

Cho, S.-J., Brown-Schmidt, S., & Lee, W.-y. (2018). Autoregressive generalized linear mixed effect models with crossed random effects: An application to intensive binary time series eye-tracking data. *Psychometrika*, *83*, 751–771. https://doi.org/10.1007/s11336-018-9604-2.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*, 335–359. https://doi.org/10.1016/s0022-5371(73)80014-3.

Cox, M. D. (1970). *The analysis of binary data*. London: Chapman and Hall.

Craigmile, P. F., Peruggia, M., & Van Zandt, T. (2010). Detrending response time series. In S.-M. Chow, E. Ferrer, & F. Hsieh (Eds.), *Statistical methods for modeling human dynamics: An interdisciplinary dialogue* (pp. 213–240). Boca Raton, FL: Taylor & Francis.

Curran, P. J., & Bollen, K. A. (2001). The best of both worlds: Combining autoregressive and latent curve models. In L. M. Collins & A. G. Sayer (Eds.), *Decade of behavior. New methods for the analysis of change*. Washington, DC:

American Psychological Association. https://doi.org/10.1037/10409-000.

Curran, P. J., Lee, T., Howard, A. L., Lane, S., & MacCallum, R. (2012). Disaggregating within-person and between-person effects in multilevel and structural equation growth models. In J. R. Harring & G. R. Hancock (Eds.), *CILVR series on latent variable methodology. Advances in longitudinal methods in the social and behavioral sciences* (pp. 217–253). Charlotte, NC: IAP Information Age Publishing.

De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, *48*, 1–28. https://doi.org/10.18637/jss.v048.c01.

De Boeck, P., & Cho, S.-J. (2019). *IRTree modeling of cognitive processes based on outcome and intermediate data*. College Park: Maryland Assessment Research Center (MARC).

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence-Erlbaum.

Fahrmeir, L. (1992). Posterior mode estimation by extended kalman filtering for multivariate dynamic generalized linear models. *Journal of the American Statistical Association*, *87*, 501–509. https://doi.org/10.2307/2290283.

Gamerman, D. (1998). Markov chain Monte Carlo for dynamic generalised linear models. *Biometrika*, *85*, 215–227. https://doi.org/10.1093/biomet/85.1.215.

Greven, S., & Kneib, T. (2010). On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika*, *97*, 773–789. https://doi.org/10.1093/biomet/asq042.

Hallett, P. E. (1986). Eye movements. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance* (Vol. 1, pp. 10.1–10.112). New York, NY: Wiley.

Hamaker, E. L., & Grasman, R. P. (2015). To center or not to center? Investigating inertia with a multilevel autoregressive model. *Frontiers in Psychology*, *5*, 1492. https://doi.org/10.3389/fpsyg.2014.01492.

Hanna, J. E., 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. https://doi.org/10.1016/s0749-596x(03)00022-6.

Hsiao, C. (2003). *Analysis of panel data*. New York: Cambridge University Press. https://doi.org/10.1017/CBO9781139839327.

Hung, Y., Zarnitsyna, V., Zhang, Y., Zhu, C., & Wu, C. J. (2008). Binary time series modeling with application to adhesion frequency experiments. *Journal of the American Statistical Association*, *103*, 1248–1259. https://doi.org/10.1198/016214508000000508.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446. https://doi.org/10.1016/j.jml.2007.11.007.

Jahng, S., Wood, P. K., & Trull, T. J. (2008). Analysis of affective instability in ecological momentary assessment: Indices using successive difference and group comparison via multilevel modeling. *Psychological Methods*, *13*, 354–75. https://doi.org/10.1037/a0014173.

Jeon, M., & De Boeck, P. (2016). A generalized item response tree model for psychological assessments. *Behavior Research Methods*, *48*, 1070–1085. https://doi.org/10.3758/s13428-015-0631-y.

Jeon, M., & Rabe-Hesketh, S. (2016). An autoregressive growth model for longitudinal item analysis. *Psychometrika*, *81*, 830–850. https://doi.org/10.1007/s11336-015-9489-2.

Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika*, *75*, 70–98. https://doi.org/10.1007/s11336-009-9141-0.

Koval, P., Kuppens, P., Allen, N. B., & Sheeber, L. (2012). Getting stuck in depression: The roles of rumination and emotional inertia. *Cognition and Emotion*, *26*, 1412–1427. https://doi.org/10.1080/02699931.2012.667392.

Kuppens, P., Allen, N. B., & Sheeber, L. B. (2010). Emotional inertia and psychological adjustment. *Psychological Science*, *21*, 984–991. https://doi.org/10.1177/0956797610372634.

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, *13*, 203–229. https://doi.org/10.1037/a0012869.

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology*, *4*, 226. https://doi.org/10.3389/fpsyg.2013.00226.

Matzke, D., Dolan, C. V., Batchelder, W. H., & Wagenmakers, E.-J. (2015). Bayesian estimation of multinomial processing tree models with heterogeneity in participants and items. *Psychometrika*, *80*, 205–235. https://doi.org/10.1007/s11336-013-9374-9.

McMurray, B., Samelson, V. M., Lee, S. H., & Tomblin, J. B. (2010). Individual differences in online spoken word recognition: Implications for SLI. *Cognitive Psychology*, *60*, 1–39. https://doi.org/10.1016/j.cogpsych.2009.06.003.

McMurray, B., Klein-Packard, J., & Tomblin, J. B. (2019). A real-time mechanism underlying lexical deficits in developmental language disorder: Between-word inhibition. *Cognition*, *191*, 104000. https://doi.org/10.1016/j.cognition.2019.06.012.

Merali, N., Veeramootoo, D., & Singh, S. (2019). Eye-tracking technology in surgical training. *Journal of Investigative Surgery*, *32*, 587–593. https://doi.org/10.1080/08941939.2017.1404663.

Molenaar, P. C. (1985). A dynamic factor model for the analysis of multivariate time series. *Psychometrika*, *50*, 181–202. https://doi.org/10.1007/bf02294246.

Mozuraitis, M., Chambers, C. G., & Daneman, M. (2015). Privileged versus shared knowledge about object identity in real-time referential processing. *Cognition*, *142*, 148–165. https://doi.org/10.1016/j.cognition.2015.05.001.

Nadig, A. S., & Sedivy, J. C. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychological Science*, *13*(4), 329–336. https://doi.org/10.1111/1467-9280.00460.

Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica: Journal of the Econometric Society*, *49*, 1417–1426. https://doi.org/10.2307/1911408.

Qu, S., & Chai, J. Y. (2007). An exploration of eye gaze in spoken language processing for multimodal conversational interfaces. In C. L. Sidner, T. Schultz, M. Stone, & C. Zhai (Eds.), *NAACL HLT 2007-human language technologies 2007: The conference of the North American chapter of the association for computational linguistics, proceedings of the main conference* (pp. 284–291). Rochester, NY: The Association for Computational Linguistics.

Quené, H., & van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, *59*, 413–425. https://doi.org/10.1016/j.jml.2008.02.002.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Retrieved April 5, 2017 from http://www.R-project.org/.

Richardson, D. C., & Spivey, M. J. (2004). Eye tracking: Characteristics and methods. In G. Wnek & G. Bowlin (Eds.), *Encyclopedia of biomaterials and biomedical engineering* (pp. 1028–1042). Boca Raton: CRC Press.

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, *95*, 318–339. https://doi.org/10.1037//0033-295x.95.3.318.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592. https://doi.org/10.2307/2335739.

Rubin, R. D., Brown-Schmidt, S., Duff, M. C., Tranel, D., & Cohen, N. J. (2011). How do I remember that I know you know that I know? *Psychological Science*, *22*, 1574–1582. https://doi.org/10.1177/0956797611418245.

Ryskin, R. A., Benjamin, A. S., Tullis, J., & Brown-Schmidt, S. (2015). Perspective-taking in comprehension, production, and memory: An individual differences approach. *Journal of Experimental Psychology: General*, *144*, 898–915. https://doi.org/10.1037/xge0000093.

Salverda, A. P., Kleinschmidt, D., & Tanenhaus, M. K. (2014). Immediate effects of anticipatory coarticulation in spoken-word recognition. *Journal of Memory and Language*, *71*, 145–163. https://doi.org/10.1016/j.jml.2013.11.002.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464. https://doi.org/10.1214/aos/1176344136.

Sedivy, J. C. (2007). Implicature during real time conversation: A view from language processing research. *Philosophy Compass*, *2*, 475–496. https://doi.org/10.1111/j.1747-9991.2007.00082.x.

StataCorp. (2017). *Stata statistical software: Release 15*. College Station, TX: StataCorp LLC.

Staudte, M., & Crocker, M. W. (2011). Investigating joint attention mechanisms through spoken human–robot interaction. *Cognition*, *120*, 268–291. https://doi.org/10.1016/j.cognition.2011.05.005.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 1632–1634. https://doi.org/10.1126/science.7777863.

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, *43*, 39–55. https://doi.org/10.1111/j.2044-8317.1990.tb00925.x.

Walker, G. M., Hickok, G., & Fridriksson, J. (2018). A cognitive psychometric model for assessment of picture naming abilities in aphasia. *Psychological Assessment*, *30*, 809–826. https://doi.org/10.1037/pas0000529.

Wang, L. P., Hamaker, E., & Bergeman, C. S. (2012). Investigating inter-individual differences in short-term intra-individual variability. *Psychological Methods*, *17*, 567–581. https://doi.org/10.1037/a0029317.

West, M., Harrison, P. J., & Migon, H. S. (1985). Dynamic generalized linear models and Bayesian forecasting. *Journal of the American Statistical Association*, *80*, 73–83. https://doi.org/10.2307/2288042.

Manuscript Received: 29 MAY 2018
Published Online Date: 21 FEB 2020