A crossed random effects diffusion model for speeded semantic categorization decisions[☆]

Joachim Vandekerckhove[*]

Department of Public Health, University of Leuven

Steven Verheyen, Francis Tuerlinckx

Department of Psychology, University of Leuven

Abstract

Choice reaction times (RTs) are often used as a proxy measure of typicality in semantic categorization studies. However, other item properties have been linked to choice RTs as well. We apply a tailored process model of choice RT to a speeded semantic categorization task in order to deconfound different sources of variability in RT. Our model is based on a diffusion model of choice RT, extended to include crossed random effects (of items and participants). This model retains the interesting process interpretation of the diffusion model's parameters, but it can be applied to choice RTs even in the case where there are few or no repeated measurements of each participant-item combination. Different aspects of the response process are then linked to different types of item properties. A typicality measure turns out to predict the rate of information uptake, while a lexicographic measure predicts the stimulus encoding time. Accessibility measures cannot reliably predict any component of the decision process.

[☆]This research was funded by K.U.Leuven Research Council grants GOA/00/02–ZKA4511, GOA/2005/04–ZKB3312, and IUAP P5/24 to FT and JV. SV is a Research Assistant of the Research Foundation – Flanders (FWO). We are indebted to Simon De Deyne for providing us with the RT data. We would like to thank Scott Brown, Michael Lee, and an anonymous reviewer for their insightful and helpful comments on an earlier draft of this article, and Roger Ratcliff for inspiring discussions on the topic. This paper is part of the doctoral project of JV.

[*]Corresponding author

Email addresses: joachim.vandekerckhove@med.kuleuven.be (Joachim Vandekerckhove), steven.verheyen@psy.kuleuven.be (Steven Verheyen), francis.tuerlinckx@psy.kuleuven.be (Francis Tuerlinckx)

Key words: semantic categorization, response times, cognitive psychometrics, hierarchical models, diffusion model

1. Introduction

In speeded semantic categorization tasks, participants are asked to verify whether a lexical item is a true member of some semantic category, and to do so as fast and as accurately as possible. Such tasks have been a primary tool in the study of semantic memory for decades. It is commonly believed that the difference in the time it takes for a participant to determine that *apple*[1] is a member of the category ***fruit*** and the time it takes for them to determine the same of *lime* may reveal important aspects of the representation of the category ***fruit*** (McCloskey and Glucksberg, 1979; Smith et al., 1974).

Historically, various views on the organization of semantic memory have succeeded one another. The types of variables that have been considered as determinants of categorization time differences offer some insight into these different views. In the original work by Landauer and Freedman (1968) and by Collins and Quillian (1970), two factors were considered important determinants of categorization time: the frequency with which lexical items appear in written discourse, and the size of the categories to which these items (supposedly) belong. In later work, researchers turned to associative accounts of the time needed to verify or discard category membership. For example, Wilkins (1971) argued that the number of times an item has been associated with the category in the past is an important determinant of the item's categorization time, while Loftus (1973) also made the argument for the importance of the reverse association. The number of times a category has been associated with an item should allow one to predict how long a participant will take to establish the set inclusion relationship between the item and the category. However, it wasn't until Wilkins' production frequency or instance dominance predictor and

[1]Throughout, we will typeset lexical entries in *italics* and categories in ***boldface italics***.

Loftus' category dominance predictor were complemented by measures of category representativeness that the speeded semantic categorization task achieved its prominence (Larochelle and Pineau, 1994). To date, the task remains best known for demonstrating that items that are representative or typical of a category are more quickly endorsed than category members that are not (Rips et al., 1973; Rosch, 1973). Since the work by Glass and Meany (1978) and by McCloskey (1980) it is now also generally recognized that whenever measures of typicality have an effect, measures of familiarity are likely to be of influence as well.

As a result of these developments, researchers who nowadays are interested in studying speeded semantic categorization decisions have no choice but to include a vast number of covariates to account for categorization time differences. This is especially true in light of the multiple methodological variations the task affords (i.e., presentation order of item and category, choice of negative instances, etc.) that prevent any single contributor to categorization time variability from emerging (Casey, 1992). The multitude of covariates that have an impact on semantic categorization time has evoked quite different attitudes towards the task. Some choose to rally against it (e.g., Kintsch, 1980), arguing that the varying findings indicate that existing accounts of the task are underspecified and lack the ability to reveal anything meaningful about the structure of semantic memory. Others see it as an opportunity to investigate the coherence and interaction among the theoretical constructs thought to underlie the various covariates. They have introduced methodological variants of the task and employed multiple regression techniques to disentangle the contributions of the covariates to the resulting categorization time differences (Casey, 1992; Chumbley, 1986; Hampton, 1997; Larochelle and Pineau, 1994; Larochelle et al., 2000).

The approach taken in the present paper is informed by both these attitudes towards the speeded semantic categorization task—we believe that current methods may be too weak, and that an in-depth investigation should account for different covariates and their interplay. In the next section, we will argue that clearly specified cognitive process models are interesting tools for extracting information from data that are known to result from processes with multiple sources of variability.

1.1. Process models and cognitive psychometrics

The statistical methods we apply in the present improve upon the classical methods in two distinct ways. Firstly, we will apply a process model that is inspired from cognitive psychology. Using a process model allows us to express the data with a concise set of parameters that have interesting psychological interpretations. Secondly, we will apply a hierarchical model in order to allow for differences between persons and between items. That item differences should not be ignored was argued very strongly by Clark (1973) and by Coleman (1964), and the detrimental effects of averaging over persons have been demonstrated by, among others, Estes (1956, 2002) and Heathcote et al. (2000). Viewing the model as a whole, each data point in the set will be conceptualized as a single realization of a specific response process, whose parameters are (at each trial) a unique combination of person-specific and item-specific parameters.

As discussed in the previous section, several item covariates have been shown to covary with semantic categorization RT to some extent. The standard methods for demonstrating these relationships have typically involved general linear models (GLMs). That is, they have focused predominantly on the mean RT (often after log-transformation). Others have focused on accuracy scores, or performed person-specific regressions (and then averaged the results). However, there have been repeated calls for extracting more of the information that is available from RTs (e.g., Heathcote et al., 1991). An alternative for this standard type of analysis is to focus rather on the response process that governs the participants' behavior (or their interactions with the items). Process modeling is very similar to usual statistical modeling in that a set of assumptions is made about regularities that are presumed to be present in the data, a set of parameters is defined that together give rise to a certain range of predicted distributions of data, and then from the empirical distributions the parameters are estimated using these predictions as a template.

For the general linear model, the assumptions are well-known: (1) the criterion is in reality a linear combination of the predictors, (2) any deviation from this pattern is noise, which follows a normal distribution with mean zero, and (3) the variance of the noise distribution is constant and independent of the predictors. These assumptions might seem quite strict, but they provide mathematical convenience and are familiar—it is quite easy to estimate the parameters of this model with readily available ('off-the-shelf') methods. In process modeling, the genesis of a model works from a different direction: assumptions about the process are made first (based on theoretical insights and prior knowledge about the world) and mathematical convenience is considered only after that. Of course, convenience decisions still come into play, but typically the plausibility of the process and the interpretability of its parameters are paramount. In the artificial category literature, process models already abound (e.g., Lamberts, 2000; Nosofsky and Palmeri, 1997), but in the domain of natural language categories they are largely unexplored.

A major advantage of this approach is that it occasionally allows us to specify different, possibly independent, components of a process that together generate the response behavior in an experimental task. In the specific case of a choice response task (like the semantic categorization task) it makes sense to assume that there is more than one factor at work in the response process at any given trial. In the model that we will use (a hierarchical diffusion model; see below), separate parameters are included for a person's ability in the task (i.e., their propensity to give a correct response, irrespective of the item properties), but also of their carefulness and the speed with which they are able to execute a motor response—all parameters that can reasonably be taken to influence the eventual RT. Additionally, there are separate parameters for the degree to which an item evokes a correct[2] response, and how long it takes to encode it before

[2]It should be noted here that the accuracy of a categorization response can be somewhat subjective. For example, is a *tomato* a ***vegetable*** or a ***fruit***? Is a *raft* a ***vehicle***? For the

a semantic decision is made. This allows for an analysis with a level of detail that is not possible with conventional methods like the general linear model.

When applying such a cognitive process model, we will explicitly allow for individual differences (both between participants and between items) by embedding the model in a *hierarchical* structure. We will in other words assume that while individuals (or items) are not identical in their cognitive process parameters, they are all members of some superordinate population. In this way, hierarchical models are a compromise between assuming that all participants are interchangeable (Batchelder, 2007) and can hence be averaged over (possibly leading to averaging artifacts; Estes, 1956, 2002; Heathcote et al., 2000), and assuming that they share no commonality at all. This hierarchical structure is a second (but equally important) way in which our method improves upon the traditional approach. As an additional feature of hierarchical models, we will be able to (attempt to) explain part of the observed variance in parameters, through the use of covariate information (De Boeck and Wilson, 2004). Using a process model in this fashion is sometimes called *cognitive psychometrics* (Batchelder, 2007; Batchelder and Riefer, 1999).

1.2. Paper outline

The outline of the rest of the paper is as follows. In the next section, we will briefly describe one data set (due to De Deyne, 2008) that contains speeded semantic categorization data. Then we describe the so-called *Leuven Natural Concept Database* (LNCD; De Deyne et al., 2008) which contains many possible covariates of the categorization time differences observed by De Deyne (2008). Then we will analyze this joint data set using the classical approach and one extension of it (multiple linear regression with crossed random effects), but the results will turn out to be inconsistent and unclear. In the section after that,

purposes of the present paper "true category membership" was determined a-priori by the experimenters, but was kept uniform across the different data sets (see section Data sets for more details).

we will describe the *hierarchical diffusion models* (HDM) statistical framework (Vandekerckhove et al., 2009) which we believe is well suited for the analysis of this coupled data set. Then we perform this analysis and discuss the results. We conclude with a discussion of the application of process models for the purpose of disentangling different sources of variability in choice RTs and implications for semantic categorization studies.

2. Data sets

2.1. Speeded semantic categorization data

The semantic categorization data are due to De Deyne (2008). The participants were eight male and thirty-six female students of the University of Leuven, who were paid the equivalent of $10 per hour for their participation.

Each of these participants provided speeded semantic categorization decisions for each of eight categories (***birds***, ***fish***, ***insects***, ***mammals***, ***musical instruments***, ***reptiles***, ***tools***, and ***vehicles***). All items that were listed as exemplars of these categories in the LNCD served as targets in the experiment. An exemplar generation task that was described in Ruts et al. (2004) informed the construction of these lists. This resulted in the inclusion of some items that could not be considered true category members (e.g., *dolphin* as an exemplar of ***fish***). De Deyne (2008) decided not to retain these items as targets for his semantic categorization experiment. In addition, he excluded all items that were composed of more than one word (e.g., *adjustable spanner*). For each category the resulting targets were complemented by an equal number of distractors. For the natural kind categories (***birds***, ***fish***, ***insects***, ***mammals***, and ***reptiles***) related items from the domain of animals constituted the distractors (e.g., *platypus*, *lobster*, *amoeba*, *seahorse*, and *octopus* for the respective categories). For the artefact categories (***musical instruments***, ***tools***, and ***vehicles***) related artefacts served the part (e.g., *microphone*, *camera*, and *container* for the respective categories).

All participants provided categorization decisions for all items. Instructions stressed both speed and accuracy. Following a recommendation by Hampton (1997), De Deyne (2008) opted for a blocked presentation order of categories. At the onset of a block, participants were informed about the category that would have to be referenced by displaying the category label for 3500ms on the screen. Those targets and distractors that were assigned to that particular category were then presented one by one in a randomized order. Each trial consisted of the presentation of a mask (500ms), a fixation point (500ms), a blank (500ms), and the stimulus word. The stimulus word was presented for a maximum of 1800ms or until the participant responded by pressing one of two buttons on a response-box. A blank screen (800ms) separated consecutive trials. Presenting the items one at a time, blocked per category, should remove the random variance in RT that would occur if a new category label had to be read on each trial.

Participants were familiarized with the procedure through the completion of a practice block. They then completed the experimental blocks in a randomized order.

2.2. The Leuven Natural Concept Database

The introduction to the semantic categorization task already provided a brief overview of the variables that have been found to have an impact on participants' performance. They are of a diverse nature, including measures that pertain to semantic categories' internal structure (e.g., Typicality), the availability of the categories' items (e.g., Word Frequency and Familiarity) and the co-occurrence of category label and category items in the categories' learning history (e.g., Category Dominance and Production Frequency). In order to disentangle the contributions of these variables to task performance it is crucial that they are collected within a homogeneous population, since cultural or regional differences are known to affect the pattern of intercorrelations (Hampton and Gardiner, 1983; Larochelle and Pineau, 1994). The data in the LNCD (De Deyne et al., 2008) meet this condition, as all norms were collected within a few years' time with University of Leuven students. They are therefore well suited for the endeavor at hand: the semantic categorization data collection by De Deyne (2008) took place in the same student population that provided the normative data and all target category members were selected from the LNCD. Hence, the available data allow an investigation of the differences that arise among true category members in speeded semantic categorization.

Following Hampton (1997), we focused on five covariates to account for these differences: Typicality, Familiarity, Word Frequency, Production Frequency, and Word Length. All five variables are included in the LNCD and below we will briefly remind the reader how they were collected. Although we agree with Hampton that these variables are generally of interest in the speeded semantic categorization literature, the choice for this set of covariates should not be taken to imply a strong theoretical commitment by the authors. Had the LNCD included a measure of category dominance, for instance, then we would have included it in our analyses. Nor should the absence of variables like imageability or age of acquisition in our analyses be considered as a stance against their role in semantic processing. Our choice for the named five variables merely reflects the aspiration of demonstrating an approach that we believe to be valuable, using a set of theoretically justified variables.

2.2.1. Typicality (T)

The representativeness of a category's items can be assessed in a variety of ways. One of them requires participants to indicate on a Likert-type scale how typical each category item is of the category. Students who provided typicality ratings for the LNCD, indicated on a scale ranging from 1 to 20 how typical they found each category member to be (De Deyne et al., 2008).

2.2.2. Familiarity (F)

Familiarity was assessed in a similar way. Participating students stepped through a list of category items, indicating on a five point Likert-type scale how familiar they were with each of the items. A rating of 1 indicated that they had never seen, heard, or used the word before. A score of 2 indicated that they had seen, heard, or used the word just once or twice. A score of 3 indicated that they had sometimes seen, heard, or used the word. A score of 4 indicated that they had seen, heard, or used the word often. A score of 5, finally, indicated that they had seen, heard, or used the word very often.

2.2.3. *Word frequency (W)*

A measure of item availability that is not based on participants' judgements, but on the frequency with which the item appears in written discourse, can also be obtained from the LNCD (see also Steyvers, this issue). The reported word frequencies in De Deyne et al. (2008) are the logarithmically transformed lemma counts taken from the Dutch CELEX lexical database (Baayen et al., 1993).

2.2.4. *Production frequency (P)*

For each of the category members, the measure of production frequency that is distributed with the LNCD tallies how many out of a total of 120 student participants generated the member in response to the category label. For the purposes of all following analyses, the production frequencies were incremented by one and logarithmically transformed.

2.2.5. *Word length (L)*

The variable word length finally, contains the number of characters in each category member. The effect of this lexicographic variable is usually of minor theoretical importance in accounts of semantic categorization and therefore regularly overlooked (imprudently, our results suggest).

2.2.6. *Covariate preprocessing*

Each of the covariates was transformed to a standardized scale with mean 0 and standard deviation 0.1. The distractor items (i.e., items that were not true members of the target category) were included in the analysis after the standardization (i.e., their covariate scores were not used to compute the standard deviation of the covariate). For most covariates, we had no information regarding these distractors, and they were assigned a value of 0 accordingly. Only for the variable Word Length (which is of course easy to obtain) were the distractors given a value.

3. Regression analysis

We subjected the joint data set to a multiple linear regression. For the present analysis, we removed all error responses and all responses that were faster than 250ms or slower than 1800ms (which was the experimental cut-off). Using the logarithm of RT as the criterion variable, and Typicality T, Length L, Familiarity F, Word Frequency W, Production Frequency P, and category membership C as predictors, the following regression model is obtained:

$$\begin{aligned} \mu_{(i)} &= \beta_0 + \beta_1 T_{(i)} + \beta_2 L_{(i)} + \beta_3 F_{(i)} + \beta_4 W_{(i)} + \beta_5 P_{(i)} + \beta_6 C_{(i)} \\ \log(RT_{(pi)}) &\sim N(\mu_{(i)}, \sigma^2). \end{aligned}$$

In this model, $RT_{(pi)}$ is the RT of person p ($p = 1, \ldots, 45$) to item i ($i, \ldots, I$), $\mu_{(i)}$ is the predicted value of $\log(RT_{(pi)})$, and σ^2 is the unexplained variance. Note again that since we only have covariate information for target items, all covariates except Word Length L and Category Membership C take the value 0 for all distractor items.

The regression results are summarized in Table 1. We have immediately performed inference on these results, and omitted all regression weights that turned out to be not statistically significant. In this way, the table concisely portrays the conclusions that would usually be drawn from the data with respect to sign and significance. Unfortunately, inspection of Table 1 shows that the picture is inconsistent with the results found in the literature, where the Typicality measure was traditionally found to have a negative effect on RT (i.e., higher Typicality leads to shorter RTs; Rips et al., 1973; Rosch, 1973). In the present data set this effect only surfaces in three out of eight categories. In two categories RT increased with Typicality, and in the remaining three categories, no effect can be discerned. The Length measure is the only one with effects that are consistent across categories, but it only significantly increased RT in three categories. For Familiarity, Word Frequency, and Production Frequency, the regression weights can take either sign, depending on categories.

Retaining error responses or not removing fast and slow responses affected the pattern of significance, but in no case did a consistent pattern arise. Hence, the classical analysis yields disappointing results.

Table 1: Classical linear regression. The signs of the regression weights whose p-value was less then 0.05 are displayed, others are replaced by a dot.

Category	T	L	F	W	P
birds	·	·	·	·	−
fish	·	·	+	+	·
insects	+	+	−	·	·
mammals	−	·	·	·	+
musical instruments	−	+	+	−	+
reptiles	+	+	·	+	·
tools	·	·	·	·	+
vehicles	−	·	+	·	·

4. Regression analysis with crossed random effects

In a second analysis, we include random effects in the regression analysis. We include this extension of the typical multiple regression analysis in order to focus our comparison on the application of a process model, rather than on our addition of random effects. We used the same data preprocessing as in the previous section, and obtain the following model:

$$\begin{aligned}
\mu_{(i)} &= \beta_0 + \beta_1 T_{(i)} + \beta_2 L_{(i)} + \beta_3 F_{(i)} + \beta_4 W_{(i)} \\
&\quad + \beta_5 P_{(i)} + \beta_6 C_{(i)} \\
\psi_{(i)} &\sim N(\mu_{(i)}, \sigma^2_{\psi}) \\
\chi_{(p)} &\sim N(0, \sigma^2_{\chi}) \\
\log(RT_{(pi)}) &\sim N(\psi_{(i)} + \chi_{(p)}, \sigma^2).
\end{aligned}$$

Table 2: Regression weights in the crossed-random effects regression model. The signs of the regression weights whose 95% credibility intervals do not contain 0 are displayed, others are replaced by a dot.

Category	T	L	F	W	P
insects	·	·	·	·	·
musical instruments	·	+	·	·	·
reptiles	·	·	·	·	·
fish	·	·	·	·	−
vehicles	−	·	·	−	·
birds	·	·	·	·	·
tools	·	·	·	·	·
mammals	−	·	·	·	·

Here, $\psi_{(i)}$ is the unique contribution of item i, while $\chi_{(p)}$ is that of person p. σ^2_ψ and σ^2_χ denote, respectively, the variability between items and between persons. Averaged over semantic categories, $\sigma_\psi = 0.0631$ and $\sigma_\chi = 0.1252$.

The regression results are again summarized in Table 2, in the same way as in the previous section. (With a difference being that we performed this analysis in a Bayesian statistical framework and we no longer apply p-values, but 95% credibility intervals instead.) Results of this second analysis are more consistent than those of the previous, but dishearteningly few regression weights turn out to differ from zero.

Again, changes in the preprocessing of the data do not meaningfully alter the results. As in the previous section, the linear model yields disappointing results. In the next section, we introduce a process model for choice RT with which we will reanalyze the present data.

5. Hierarchical diffusion models

In the domain of choice RTs, models based on the Wiener diffusion process have garnered significant attention (Ratcliff, 1978; Ratcliff and Rouder, 1998; Ratcliff and Smith, 2004). The Wiener diffusion model is one of the broad class of sequential sampling models where, in this case, a single evidence counter evolves over continuous time until it hits one of two absorbing boundaries. The time to absorption is then related to the RT and which boundary was hit indicates the response given. The model is considered particularly interesting because the parameters that drive the process (explained below) have intuitive interpretations relating to the sequential accumulation of information.

The increasing popularity of the diffusion model for choice RTs is likely due in part to the interesting interpretations of its parameters on the one hand, and the model's ability to account for many empirically observed phenomena on the other (for an excellent review of recent advances with the diffusion model, see Wagenmakers, in press). It is unfortunate, therefore, that the possibilities for application of the diffusion model have thus far been somewhat limited. For example, as noted by Wagenmakers (in press), fitting the diffusion model to empirical data requires a large number of observations. Importantly, with the methods currently in practice (Ratcliff and Tuerlinckx, 2002; Vandekerckhove and Tuerlinckx, 2007, 2008; Voss and Voss, 2007) it has typically been necessary to have an appreciable number of data points in each cell of the experimental design. That is, some independent replications *under invariant conditions* are required in order to obtain parameter estimates.

As a result, applications of the diffusion model have largely been limited to the analysis of "long" data sets (i.e., a typical psychophysical design where there are few participants, and many repeated trials for each participant and in each condition). A little-explored alternative would be to apply it to "wide" data sets with many participants and few repeated measurements (like the present semantic categorization data set; Hampton, 1997, recommends against repeating items in such a paradigm). Such analyses are more challenging for several reasons. For example, if all participants are analyzed independently of one another, the available number of data points on which the estimates would be based would be very low. On the other hand, it would be unreasonable to keep many parameters constant across individuals, making it impossible to pool the data together (i.e., to allow sharing of information between data from different participants). Other methods of pooling data across participants (or, for that matter, items), such as quantile averaging (or *vincentizing*; Ratcliff, 1979; Rouder and Speckman, 2004), come with preconditions that may not be met by the diffusion model (i.e., same location-scale family; Thomas and Ross, 1980), they do not permit an investigation of individual differences (in which we might be interested), and they cannot be applied in the case where individual differences are expected on both the person and the item side (i.e., if both persons and items are random draws from their respective populations, and there are no repetitions of person-by-item combination trials, then there are no distributions left to combine). Taking the statistically principled route of treating participants as random samples from a population (random effects approach to individual differences) typically leads to models that rapidly become quite complex mathematically.

Ratcliff (1978), Laming (1968), and Link and Heath (1975) have already approached part of this problem with a random-effects strategy, by allowing *trial-to-trial variability* in model parameters. Effectively, it is assumed that some parameters change over time in that they are, at each point in time, a random sample from some higher-order distribution. Parameters of this superordinate distribution are then estimated in lieu of the trial-specific parameters themselves. In practice, the variability in a parameter is implemented by multiplying the model's likelihood function with the likelihood function of the trial-to-trial variability and then integrating over the parameter(s) that is (are) allowed to vary (see Ratcliff and Tuerlinckx, 2002; Tuerlinckx, 2004). However, this method is computationally laborious (involving multidimensional integrations that have to be approximated by sums) and somewhat inflexible (in the sense that the likelihood function has to be adapted in such a way that makes it difficult to write a universal but efficient algorithm).

Recently, Vandekerckhove et al. (2009) have approached this problem by applying the flexibility of Bayesian hierarchical modeling (see, e.g. Rouder and Lu, 2005; Rouder et al., 2005, 2007) for some examples of Bayesian hierarchical modeling) to the Wiener diffusion process. This statistical framework (HDM) can easily cope with many simultaneous random effects, and software for its implementation is freely available (Vandekerckhove et al., 2009). A diffusion model with crossed random effects can be applied to a data set where there are no repeated observations in the item-by-participant cells of the experimental design. Such a design would be inaccessible to typical process model approaches, but it is important in order to account for the random sampling scheme that is normally used for lexical items in the semantic categorization paradigm (Clark, 1973; Coleman, 1964) as well as participants. This crossed random effects diffusion model is especially suited for this case, because other methods that are typically used for dealing with uncontrolled variability (e.g., vincentizing) cannot cope with the crossed random effects design without repeated observations of each cell of the design. Additionally, even if it were possible to have repeated observations of the same person/item combinations (this is not recommended in the semantic categorization context according to Hampton, 1997, but it might be possible in other cases), then the vincentization procedure would only allow us to account for the variability. It does not permit an easy quantification of the variability, nor would it allow us to attempt to explain the variability through external covariates (De Boeck and Wilson, 2004).

5.1. Diffusion models

At the basis of the Wiener diffusion model is a random walk process in continuous time and with a continuous state space. The random walk has two boundaries at values α ("upper") and 0 ("lower") and its step-size over a discrete time period t is a randomly drawn value from $N\left(\delta t, \sigma^2 t\right)$ (Feller, 1970). By convention, $\sigma = 0.1$. After a number of steps, the process will hit one of its boundaries (see Fig. 1). If δ, called the *drift rate* (or *drift* for short), is high in absolute value, then the number of steps will be small. The boundary that was hit is then linked to the response given, and the first passage time (i.e., the number of steps taken to reach the boundary) relates to the RT. By convention, a hit at the upper boundary (α) is linked to correct responses and a hit at the lower boundary (0) is an error.[3] Of prime interest in the modeling of choice reaction times are the proportions with which the absorbing boundaries are hit, as well as the predicted first passage time distribution at either boundary. The two parameters of this model (sometimes also called the *drift diffusion model*) have straightforward interpretations. Boundary separation α relates to the amount of information that is required to make a decision—that is, it indicates the caution level of the decision system (in this case the participant). We will therefore usually let α be different for different persons (but identical within experimental blocks, because we do not expect people to adapt their caution level in the middle of an experimental block). The second parameter, drift rate δ, is the speed of information accumulation. We can easily suppose this to depend both on the participant (who may be more or less able to rapidly process information) and on the item (which may be relatively rich or poor in information content).[4]

Typically two more parameters are introduced to the unbiased drift diffusion model. Firstly, a *bias* parameter to indicate that the starting point of the

[3]This convention can be adapted. We could for example say that the upper boundary indicates a category affirmation response and the lower is connected to a category negation. However, the interpretation of the parameters would then change: a high drift rate would no longer evoke a correct response, but rather the category affirmation response, whether correct or not.

[4]In principle, it would also be possible that some participants have a better affinity with some items, resulting in a person-by-item interaction. Such an interaction could be most interesting when, for example, comparing groups of participants with different levels of experience with a certain semantic category (e.g., comparing ichthyologists with laymen in their categorization performance of ***fish***). However, including a person-specific interaction would require more than one replication of each person-by-item combination, which the present data set does not offer.

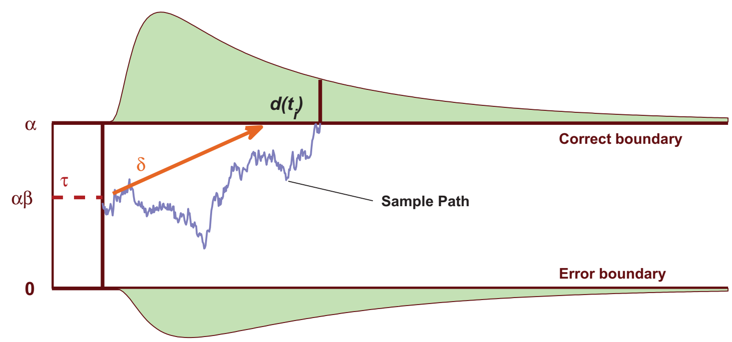

Figure 1: A graphical illustration of the Wiener diffusion model.

process may be closer to 0 or to α. This parameter is $\beta \in [0,\ 1]$, so that the starting value of the process is exactly $\alpha\beta$. Secondly, a shift parameter τ is added to represent RT components that are not part of the decision time (e.g., encoding the stimulus and executing the motor response). The nondecision time is assumed to be stochastically independent from the decision time. The joint probability density of the RT and accuracy (i.e., the Wiener likelihood function, or its probability density function, PDF) is given in Tuerlinckx (2004), and we denote it with $W_{X,T}(x, t|\alpha, \tau, \beta, \delta)$, where the random variables X and T represent the response given and the response latency, respectively. Instances of X and T will be denoted as x and t.

5.2. Hierarchical extension

In a hierarchical diffusion model (HDM; Vandekerckhove et al., 2009), the four parameters that drive the response process are considered random draws from some partly specified distribution (Rouder et al., 2005) that may be subject to many different influences. For example, it may be assumed that the drift rate $\delta_{(i)}$ of the response process at trial i is a random draw from a normal distribution with mean ν and standard deviation η:

$$\delta_{(i)} \sim N\left(\nu, \eta^2\right).$$

The parameters of this distribution can in turn be considered random draws from some higher-order distribution, or they may be seen as some fixed function of other parameters or of data. The multitude of combinations that are possible make the HDM framework an exceedingly flexible method for the analysis of two-choice RT data.

5.3. Bayesian implementation

Obtaining parameter estimates in such a flexible context would be quite challenging in general. Finding the maximum-likelihood parameter estimates for a random-effects diffusion model would require repeated computations of a multidimensional integral over the (already nontrivial) Wiener distribution. However, the inclusion of randomly varying parameters and integrating over their distributions is the basic modus operandi in Bayesian statistics. Hence, casting the HDM in a Bayesian statistical framework (bulding upon Vandekerckhove et al., 2008) allows us to apply the model easily. In the next section, we will specify a specially-tailored HDM, which we will then apply to the semantic recognition data.

6. Analyzing the semantic recognition data

6.1. Model assumptions

Formally speaking, a statistical model is little more than a set of assumptions regarding structure that is present in the data. We discern five different types of assumptions in the present model, which we describe in turn. We will apply this model to each category separately.

6.1.1. The measurement model

At the most basic level, our assumption is that each data point is generated by a diffusion process whose parameters may differ between persons and/or items (i.e., words). We have chosen to allow boundary separation α to differ between persons p, while nondecision time τ and drift rate δ may be different for each item-by-person combination pi. Since we do not want to assume that participants have an a-priori bias for the correct or erroneous responses[5], we fix the bias β to 0.5 for the remainder of this presentation. Formally, the measurement model is written as follows:

$$(t_{(pi)}, x_{(pi)}) \quad \sim \quad W(\alpha_{(p)}, \beta, \tau_{(pi)}, \delta_{(pi)}).$$

This gives the expected distribution of data point $(t_{(pi)}, x_{(pi)})$ (for person p on item i) given all the relevant parameters. W is the Wiener PDF. Note that, as mentioned in an earlier section, we do not let boundary separation α depend on items, so it does not receive an index i.

6.1.2. Trial-to-trial variability

Parameters τ and δ are assumed to vary both between persons and between items (and hence from trial to trial). For this random variability, we assume a normal distribution, which is the most common assumption in hierarchical modeling (e.g., De Boeck and Wilson, 2004) and we see no reason to deviate from it here.[6] The normal also serves as a useful first approximation. In both cases, we allow the mean of the trial-to-trial distribution to depend on both persons and items. The dependence on persons is simply to allow for interindividual differences (which we believe exist), but the dependence on items is crucial in order to explain interitem differences with the LNCD covariates. Formally:

$$\begin{aligned} \delta_{(pi)} &\sim N(\nu_{(pi)}, \eta^2_{(p)}) \\ \tau_{(pi)} &\sim N(\theta_{(pi)}, \phi^2_{(p)}). \end{aligned}$$

---

[5] We can safely assume this, since there were 50% targets and 50% distractors in each block of the experiment.

[6] In principle, one could object that τ cannot follow a normal distribution, as it can never be negative, but in practice the mean (θ) of this distribution has always turned out to be very large compared to its standard deviation (ϕ), so that the mass of this distribution below zero can be safely ignored.

It can be seen that we also allow for the possibility of different trial-to-trial variances between persons.

6.1.3. Independent item and person contributions

As explained in the previous section, we want items and persons to have independent effects on two different aspects of the decision process. For the drift rate $\delta_{(pi)}$, we call the item and participant contributions $\lambda_{(i)}$ and $\gamma_{(p)}$, respectively. For the nondecision time $\tau_{(pi)}$ we call them $\psi_{(i)}$ and $\chi_{(p)}$. In both cases, we assume these to be independent and additive (this is a typical construction in psychometrics; for example the Rasch model uses the same assumption; De Boeck and Wilson, 2004):

$$\begin{aligned}\nu_{(pi)} &= \gamma_{(p)} + \lambda_{(i)} \\ \theta_{(pi)} &= \chi_{(p)} + \psi_{(i)}.\end{aligned}$$

6.1.4. Population distributions

Since both items and participants were random samples from a larger population, a random effects design is appropriate. Those parameters that have a population distribution are thus assigned population-level parameters. Two distributions over the item population must be defined: that of the item contribution to the drift rate (i.e., $\lambda_{(i)}$) and of the item contribution to the nondecision time (i.e., $\psi_{(i)}$). These components again get normal population distributions:

$$\begin{aligned}\lambda_{(i)} &\sim N(\mu_{\lambda(i)}, \sigma^2_{\lambda(i)}) \\ \psi_{(i)} &\sim N(\mu_{\psi(i)}, \sigma^2_{\psi}).\end{aligned}$$

Note that, since we expect the drift rates for targets and items to be quite different, we also allow their population variances to be different.

For reasons of model identifiability, the mean of one of the random components must be constrained, so we set the mean of $\gamma_{(p)}$ and $\chi_{(p)}$ to 0:

$$\begin{aligned}\gamma_{(p)} &\sim N(0, \sigma^2_{\gamma}) \\ \chi_{(p)} &\sim N(0, \sigma^2_{\chi}).\end{aligned}$$

Finally, we define a population distribution for the boundary separation α:

$$\alpha_{(p)} \sim N(\mu_\alpha, \sigma_\alpha^2).$$

6.1.5. Regression structure

We make most of the previous assumptions in order to account for the possibility of variation between persons or items. Until now, however, the model is strictly descriptive (i.e., it does not include any external covariates that might be employed to explain the variability that is observed). A final set of assumptions pertains to the relationship between the diffusion model parameters and the LNCD. Following Hampton (1997), we include five covariates: Typicality (T), Word Length (L), Familiarity (F), Word Frequency (W), and Production Frequency (P). All of these covariates were standardized to have a mean of 0 and a standard deviation of 0.1. As in the regression analysis we showed before, we also add the item's category membership as a predictor (i.e., $C_{(i)} = 1$ if the item was a target, $C_{(i)} = 0$ if it was a distractor). We call the regression weights ζ for the drift rate and ρ for the nondecision time:

$$\begin{aligned}
\mu_{\lambda(i)} &= \zeta_0 + \zeta_1 T_{(i)} + \zeta_2 L_{(i)} + \zeta_3 F_{(i)} + \zeta_4 W_{(i)} + \zeta_5 P_{(i)} + \zeta_6 C_{(i)} \\
\mu_{\psi(i)} &= \rho_0 + \rho_1 T_{(i)} + \rho_2 L_{(i)} + \rho_3 F_{(i)} + \rho_4 W_{(i)} + \rho_5 P_{(i)} + \rho_6 C_{(i)}.
\end{aligned}$$

In principle, we could do the same for the person contributions to the drift rate or nondecision time, or for the caution parameter α. For example, intelligence might predict the drift rate component (see e.g., Ratcliff et al., 2008) or neuroticism might be connected to the caution parameter. Unfortunately, the present data sets do not include person covariates.

6.2. Results

The model we have presented is an instance of a hierarchical diffusion model. Software to implement such a model was made available by Vandekerckhove et al. (2009, "wienereta.odc"). Using this software, we obtained posterior distributions for each of the parameters in the model.[7] The Appendix contains a discussion of the fit of the model to the data.

The posterior distributions for ζ, per semantic category, are displayed in Fig. 2, and those for ρ are in Fig. 3. These posterior inference plots may be read as follows. In each of the subplots, the five horizontal lines represent the posterior distributions of the regression weights of the five LNCD covariates. The lines indicate the Bayesian credibility interval (CI): the region around the mean that contains 95% of the mass of the parameter's posterior distribution. The diamonds indicate the posterior means. The vertical line is the value 0. In these figures, two patterns emerge quite clearly: the effect of Typicality (T) on drift rate is always positive, and most of those CIs do not include 0. Similarly, in Fig. 3, Word Length (L) generally has a positive effect on nondecision time.

To compare these results to the ones obtained from the standard analysis in Table 1, we constructed a similar table for these two sets of regression weights. We display the sign of a regression weight if its 95% credibility interval does not contain 0 (i.e., with 95% probability the parameter is not 0). In contrast with the classical analysis, results here are predominantly consistent—for the drift rate regression, only Typicality consistently shows up as a good predictor. For the nondecision time, Word Length has a consistent influence. In both cases, the sign of the regression weight is as expected.[8]

Fig. 4 shows the relationship between the Typicality score of an item and its contribution to the drift rate (depicted for an average participant; i.e. $\gamma_{(p)} = 0$) in the category ***mammals***. A somewhat linear relationship is evident[9], and we

[7] We followed the recommendations made by Vandekerckhove et al. (2009) to check for convergence issues and found that there were none (all convergence statistics $\hat{R} < 1.05$, all chains show proper mixing).

[8] The present analysis is based on a multiple regression. In one alternative attempt, we restricted ourselves to univariate regressions (i.e., including one covariate at a time), and obtained comparable results.

[9] The linear relationship is clearly not perfect, and perhaps even better prediction could have been achieved with a non-linear regression, but we do not explore that avenue here.

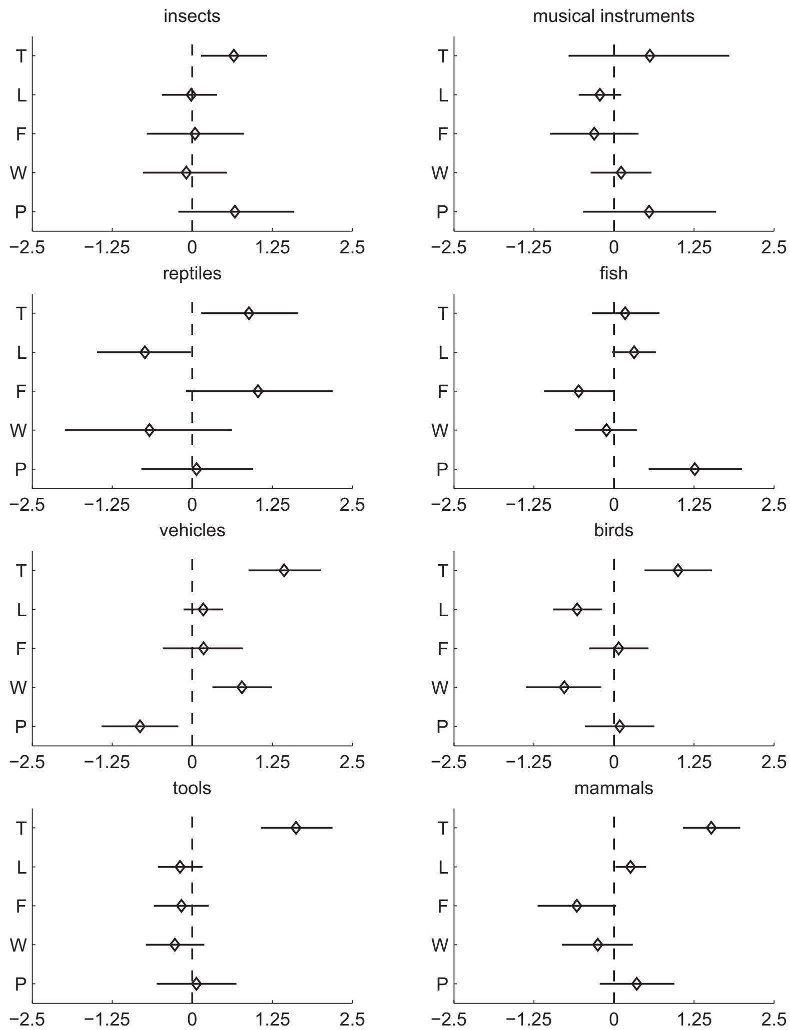

Figure 2: Posterior inference plots for the regression weights ζ (the regression weights for the λs, the item contributions to the drift rates). See text for details.

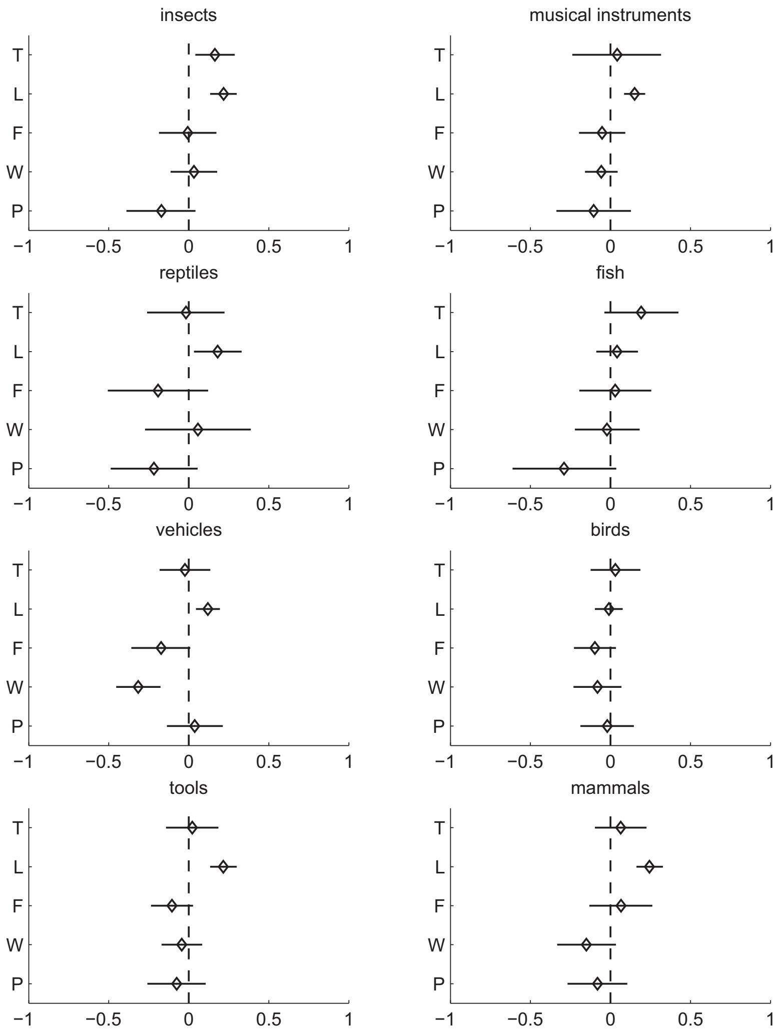

Figure 3: Posterior inference plots for the regression weights ρ (the regression weights for the ψs, the item contributions to the nondecision times). See text for details.

Table 3: Regression weights in the HDM. The signs of the regression weights whose 95% credibility intervals do not contain 0 are displayed, others are replaced by a dot.

	ζ					ρ				
	T	L	F	W	P	T	L	F	W	P
birds	+	−	·	−	·	·	·	·	·	·
fish	·	·	−	·	+	·	·	·	·	·
insects	+	·	·	·	·	·	+	·	·	·
mammals	+	+	·	·	·	·	+	·	·	·
musical instruments	·	·	·	·	·	·	·	·	·	·
reptiles	+	−	·	·	·	·	+	·	·	·
tools	+	·	·	·	·	·	+	·	·	·
vehicles	+	·	·	+	−	·	+	·	−	·

have labeled some of the items on the graph. Item *bat* has the lowest Typicality rating, and also the lowest drift rate. Items *dog* and *lion* reside on the opposite side of the spectrum.

While some of these effects are very easy to interpret, others are less intuitive. In Fig. 5, we display the effect that drift rate has on the raw data. We selected three items from the range of Typicality ratings (from the ***mammals*** category) and display the expected distribution of their (correct) RTs and their expected accuracy scores.

The interindividual variability is also notable. In particular, the person-specific α parameter that represents a person's caution shows much variation. Fig. 6 shows the effect of different boundary separations (keeping all other factors constant). We selected three participants from the population (corresponding to the 10^{th}, 50^{th}, and 90^{th} percentiles) and plotted their expected raw RT distributions and accuracies (for an average item). The range of α values in the population has a small but noticeable effect on both the RT distribution and the accuracy scores.

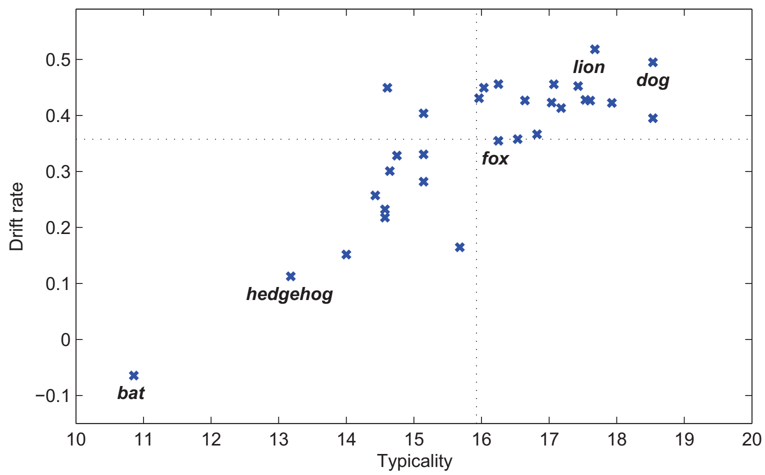

Figure 4: An example regression result. Drift rate on average increases with increasing Typicality. Item *bat* is a clear outlier on both dimensions. The dotted lines indicate the mean Typicality and mean drift rate. To avoid confusion: the values on the vertical axis are the *total* drift rates assuming an average person with $\gamma_{(p)} = 0$.

The effect of Word Length is, from a research methods point of view, perhaps the most important to keep in mind (we will elaborate on why we believe this is so in the Discussion section below). Fig. 7 shows the relationship between Word Length and nondecision time for the category ***tools*** (here, too, the effect might be better captured by a non-linear regression). The nondecision times associated with particular items range from 500ms to 630ms—the interquartile range is more than 60ms. A graphical presentation for this effect (like the ones in Figs. 5 and 6) would show identical accuracy scores and identically shaped RT distributions, but shifted to the right for items with higher values of $\psi_{(i)}$. For interpretation, we can compute that, on average, adding one letter to a word shifts the RT by 7–12ms, depending on the category.[10]

Finally, we can compare differences between participants with differences be-

[10]This is in line with results from Hampton (1997).

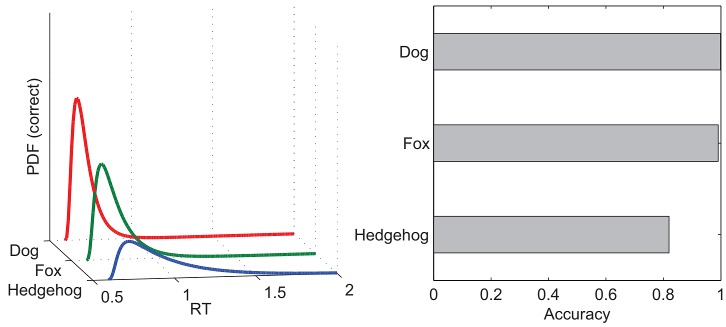

Figure 5: The effect of higher Typicality (and hence higher drift rates) on the raw data. RT distributions become more compact and less skewed (smaller mean, smaller variance) as Typicality (drift rate) increases. Accuracy increases with higher Typicality. The PDFs are marginal PDFs (i.e., conditional upon a correct answer) and have been normalized so that they integrate to 1).

tween items. Table 4 shows population standard deviations from the HDM (the values shown are the means of the posterior distributions of the parameters). Comparing the drift rate's variability due to persons (σ_γ) with its variability due to item differences ($\sigma_{\lambda(1)}$ for targets[11], $\sigma_{\lambda(2)}$ for distractors), we can see that, with the exception of the category ***fish***, the item variance is always much larger than the person variance. The reverse is true for the nondecision time: residual item variance there is much smaller than the variance in the person population.

7. Discussion

The theoretical advantages of using a process model on the one hand and a hierarchical model on the other (together leading to a cognitive psychometric model) were described in the introduction. However, the demonstration in the present article also shows the practical applicability of this method.

We believe that, as a methodological advance, the HDM framework (Van-

[11]Note that this is the *residual* item variance, after correcting for all the covariates.

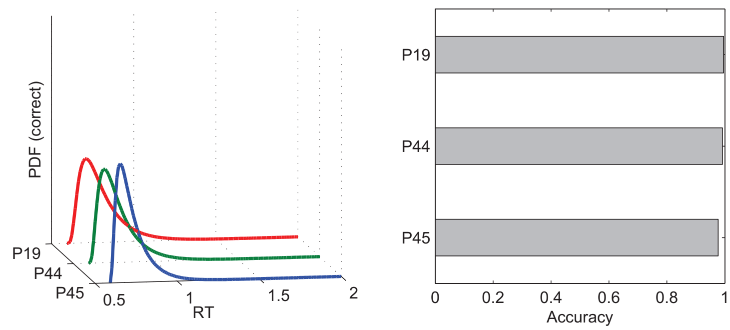

Figure 6: The effect of higher boundary separation on the raw data. P19 has a high α, P44 has a median value, and P45 has a low α. RT distributions become more skewed with increasing α, but accuracy increases. The PDFs are marginal PDFs (i.e., conditional upon a correct answer) and have been normalized so that they integrate to 1).

dekerckhove et al., 2009) can contribute not only to semantic categorization studies, but to a more general class of paradigms. If speeded binary choice RTs are collected, and if it is likely that there are interindividual (or interitem) differences, then the HDM framework might prove useful.

In the introduction, we have also referred to Estes' (1956, 2000) view on individual differences and how averaging over participants (or items) may lead to averaging artifacts. Hierarchical modeling deals with this issue in a practical and efficient way. In the domain of choice RTs, a different type of *unmodeled-variability artifact* may occur if variability in the various facets of the response process is ignored. In the particular case of the HDM, variability in the nondecision process time (i.e., encoding and processing time) can easily be confused for variability in the decision process time. Indeed, past analyses of semantic categorization data have found effects of word length on RT, but the present analysis strongly suggests this to be an artifact—word length does not predict the information uptake rate, but rather the encoding time of the process. However, accounting for this variability in nondecision time is important to achieve proper parameter estimates.

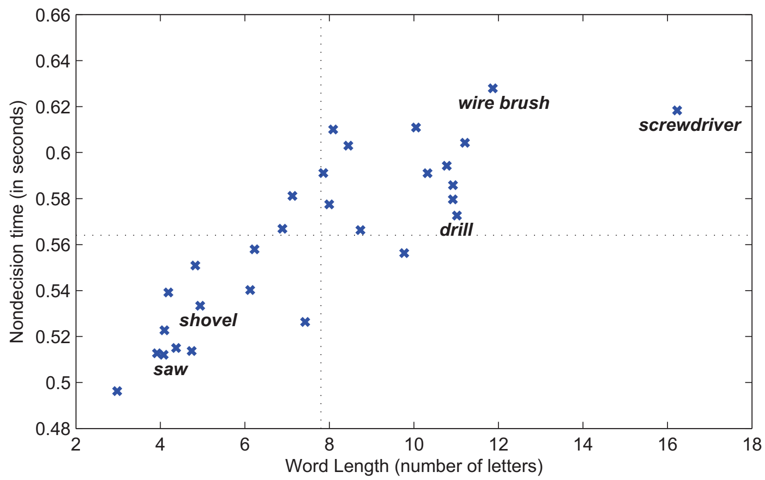

Figure 7: Nondecision time in the category ***tools*** on average increases with increasing Word Length. The dotted lines indicate the mean Word Length and mean nondecision time. The values on the vertical axis are the total nondecision times (assuming an average person with $\chi_{(p)} = 0$). Word Length has been jittered to avoid overlapping symbols. The original Dutch versions of the labeled items were (from left to right): *zaag*, *schop*, *boormachine*, *staalborstel*, and *schroevendraaier*.

7.1. Implications for semantic categorization studies: item properties

The model of speeded semantic categorization we have introduced is very explicit about the various stages involved in making a category membership decision towards a visually presented verbal stimulus. Our results suggest that elaboration of the aspects involved in arriving upon that decision is a useful practice. By attributing the effects of typicality and word length to different aspects of the response process, the analysis moves beyond the common practice of regressing these covariates upon the observed RTs. The very nature of the latter approach confines it to the mere establishment of the relative effect of both covariates upon RT. The HDM approach, by contrast, allows the effect to be attributed to specific components of the RT.

The critical reader might raise the objection that we have not been explicit enough in our account of the categorization behavior, and might point out that

Table 4: Population variability parameters in the HDM. We can compare the person variabilities with the item variabilities (see text for details). All values are standard deviations.

	boundary separation	nondecision time		drift rate		
	σ_α	σ_χ	σ_ψ	σ_γ	$\sigma_{\lambda(1)}$	$\sigma_{\lambda(2)}$
birds	0.018	0.067	0.023	0.025	0.077	0.115
fish	0.013	0.084	0.035	0.122	0.057	0.176
insects	0.022	0.064	0.024	0.018	0.113	0.229
mammals	0.024	0.067	0.023	0.019	0.069	0.085
musical instruments	0.021	0.058	0.020	0.093	0.091	0.157
reptiles	0.024	0.091	0.029	0.022	0.106	0.247
tools	0.015	0.081	0.020	0.021	0.078	0.148
vehicles	0.020	0.056	0.026	0.025	0.093	0.164

for those among us who are interested in understanding semantic cognition the question "what governs semantic categorization time differences" has shifted towards "what governs information uptake differences."

In response to this objection we readily admit that, indeed, we have been less than explicit about the representation upon which the accumulator process acts. We have not committed ourselves, for instance, to featural representations of the kind Smith et al. (1974) or McCloskey and Glucksberg (1979) have argued for. Nor did we attempt to attempt to link the accumulation process to the semantic markers that were proposed by Glass and Holyoak (1974). Although the terminology we have used throughout this manuscript (e.g., information uptake, accumulation of evidence) might tempt the reader into thinking that the diffusion model is more in favor of a successive comparison of exemplar and category features than of an ordered search through semantic markers, we do not necessarily believe this to be the case. Any representational format that allows for a stochastic accumulation of evidence for or against the endorsement of an item as a category member is in principle compatible with the diffusion model we propose (several papers in the present issue make detailed representational assumptions that could drive a sequential sampling process model: Zeigenfuse and Lee, this issue; Kemp et al., this issue; Ceulemans & Storms, this issue; Dry and Storms, this issue). This does not imply that the methodology we have proposed in this manuscript can not be brought to bear upon the representational issue. In much the same way as we have explored the relative contributions of different covariates to the degree of information uptake, one could evaluate the predictions of rivalling representations, providing that they are explicit enough to warrant quantification. One might consider using the LNCD again for such endeavors as they include plenty of information on the intension and extension of semantic categories.

For now, however, we feel that casting speeded semantic categorization decisions in terms of a diffusion model constitutes sufficient explicitation. As we have pointed out in the Introduction, much of the efforts during the last three decades have been aimed at disentangling the various constructs that are likely to influence semantic categorization. As it is along the lines of these constructs that theories of semantic behavior are likely to develop, tools that shed light on the varying manners in which they exert their influence are valuable. A HDM modeling framework may be useful in this regard.

In the near future we hope the model will allow us to study the effects of variables that are present in the LNCD, but were not incorporated in the current analyses for reasons of brevity. The question of whether age of acquisition exerts an effect in semantic categorization, and how that effect might come about, for instance, deserves some attention as they have generated considerable debate (Brysbaert et al., 2000; De Deyne and Storms, 2007; Morrison et al., 1992; Morrison and Gibbons, 2006). We also hope to study the impact category dominance has on the categorization performance participants display. This will require the collection of additional data as the LNCD does not include a direct measure of the association strength between an item and its superordinate(s).

(See De Deyne and Storms, 2008, for a discussion of the differences between the direct or constrained measures of category dominance that are mostly used in the semantic categorization literature and the unconstrained measure that can be found in the LNCD.) These and other investigations will undoubtedly benefit from experimental manipulations that are expected to influence the effect a particular covariate has on the distribution of one of the model's parameters, but not on that of others (Hampton, 1997).

7.2. Implications for semantic categorization studies: person properties

In the Results section we already indicated that our analyses implied interindividual variability in semantic categorization behavior. Namely, the person-specific α that represents a person's caution showed considerable variation with accompanying effects on the degree to which true category exemplars were endorsed as such (see Fig. 6 for a demonstration). These differences between persons reflect (more or less) imprudent task strategies resulting in (more or fewer) erroneous decisions. It has been shown (e.g., Hampton, 1998, 2007; McCloskey and Glucksberg, 1978) that people may disagree considerably about the items they consider to be true members of a semantic category. The degree to which people disagree is likely to be reflected in the variation of the α parameter.

As De Deyne (2008) did not record any information on the students participating in the semantic categorization task but their age and gender, our assumptions regarding the person side of the diffusion model have remained strictly descriptive (i.e., no external covariates that might be employed to explain the interindividual variability that was observed were available). Looking at recent applications of the speeded semantic categorization task, in which the decisions of individuals with autism were compared with those of matched controls (Gastgeb et al., 2006) or the differences in categorization behavior by Broca's and Wernicke's aphasic individuals were investigated (Kiran and Thompson, 2003), it seems that the approach argued for in this manuscript may also prove to be valuable when applied to person properties instead of item properties. One can image proposing a diffusion model of speeded semantic categorization in which person variables are regressed upon the model's parameters or a model whose parameter distributions are allowed to differ from one group to another. Along these lines we have begun to compare the categorization behavior of individuals displaying many schizotypal traits to that of individuals who display few schizotypal traits. The difference in the degree to which individuals in the general population display these traits is thought to accompany their willingness to endorse weak semantic associates as true category members (Kiang and Kutas, 2005, 2006). Hence, we would expect that in the diffusion model analysis participants scoring high on schizotypy would demonstrate a greater bias β towards the target than participants who obtained a low score.

References

References

Baayen, R. H., Piepenbrock, R., van Rijn, H., 1993. The CELEX lexical database [CD-ROM]. Philadelphia: University of Pennsylvania, Linguistic Data Consortium.

Batchelder, W. H., October 2007. Cognitive psychometrics: Combining two psychological traditions. CSCA Lecture, Amsterdam, The Netherlands.

Batchelder, W. H., Riefer, D. M., 1999. Theoretical and empirical review of multinomial process tree modeling. Psychonomic Bulletin & Review 6, 57–86.

Brysbaert, M., Van Wijnendaele, I., De Deyne, S., 2000. Age-of-acquisition effects in semantic processing tasks. Acta Psychologica 104, 215–226.

Casey, P. J., 1992. A reexamination of the roles of typicality and category dominance in verifying category membership. Journal of Experimental Psychology: Learning, Memory, and Cognition 18, 823–834.

Ceulemans, E., Storms, G., this issue. Detecting intra and inter categorical structure in semantic concepts using HICLAS. Acta Psychologica, this issue.

Chumbley, J. I., 1986. The roles of typicality, instance dominance, and category dominance in verifying category membership. Journal of Experimental Psychology: Learning, Memory, and Cognition 12, 257–267.

Clark, H., 1973. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. Learning and Verbal Behavior 12, 335–359.

Coleman, E. B., 1964. Generalizing to a language population. Psychological Reports 14, 219–222.

Collins, A. M., Quillian, M. R., 1970. Does category size affect categorization time? Journal of Verbal Learning and Verbal Behavior 8, 240–247.

De Boeck, P., Wilson, M., 2004. Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach. Springer, New York.

De Deyne, S., 2008. Proximity in semantic vector space. Ph.D. thesis, University of Leuven.

De Deyne, S., Storms, G., 2007. Age-of-acquisition differences in young and older adults affect latencies in lexical decision and semantic categorization. Acta Psychologica 124, 274–295.

De Deyne, S., Storms, G., 2008. Word associations: Norms for 1,424 Dutch words in a continuous task. Behavior Research Methods 40, 198–205.

De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M., Voorspoels, W., Storms, G., 2008. Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. Behavior Research Methods 40, 1030–1048.

Dry, M. J., Storms, G., this issue. Features of graded category structure: generalizing the family resemblance and polymorphous concept models. Acta Psychologica, this issue.

Estes, W. K., 1956. The problem of inference from curves based on group data. Psychological Bulletin 53, 134–140.

Estes, W. K., 2002. Traps in the route to models of memory and decision. Psychonomic Bulletin & Review 9, 3–25.

Feller, W., 1970. An Introduction to Probability Theory and Its Applications: Vol. I. John Wiley & Sons, New York.

Gastgeb, H. Z., Strauss, M. S., Minshew, N. J., 2006. Do individuals with autism process categories differently? The effect of typicality and development. Child Development 77, 1717–1729.

Glass, A. L., Holyoak, K. J., 1974. Alternative conceptions of semantic memory. Cognition 3, 313–339.

Glass, A. L., Meany, P. J., 1978. Evidence for two kinds of low-typical instances in a categorization task. Memory & Cognition 6, 622–628.

Hampton, J. A., 1997. Associative and similarity-based processes in categorization decisions. Memory & Cognition 25, 625–640.

Hampton, J. A., 1998. Similarity-based categorization and fuzziness of natural categories. Cognition 65, 137–165.

Hampton, J. A., 2007. Typicality, graded membership, and vagueness. Cognitive Science 31, 355–384.

Hampton, J. A., Gardiner, M. M., 1983. Measures of internal category structure: A correlational analysis of normative data. British Journal of Psychology 74, 491–516.

Heathcote, A., Brown, S., Mewhort, D. J. K., 2000. The power law repealed: The case for an exponential law of practice. Psychonomic Bulletin & Review 7, 185–207.

Heathcote, A., Popiel, S. J., Mewhort, D. J. K., 1991. Analysis of response time distributions: An example using the Stroop task. Psychological Bulletin 109, 340–347.

Kemp, C., Chang, K. K., Lombardi, L., this issue. Category and feature identification. Acta Psychologica, this issue.

Kiang, M., Kutas, M., 2005. Association of schizotypy with semantic processing differences: An event-related brain potential study. Schizophrenia Research 77, 329–342.

Kiang, M., Kutas, M., 2006. Abnormal typicality of responses on a category fluency task in schizotypy. Psychiatry Research 145, 119–126.

Kintsch, W., 1980. Semantic memory: A tutorial. In: Nickerson, R. S. (Ed.), Attention and Performance VIII. Erlbaum, Hillsdale, NJ, pp. 595–620.

Kiran, S., Thompson, C. K., 2003. Effect of typicality on online category verification of animate category exemplars in aphasia. Brain and Language 85, 441–450.

Lamberts, K., 2000. Information-accumulation theory of speeded categorization. Psychological Review 107, 227–260.

Laming, D. R. J., 1968. Information Theory of Choice-reaction Times. Academic Press, London, UK.

Landauer, T. K., Freedman, J. L., 1968. Information retrieval from long-term memory: Category size and recognition time. Journal of Verbal Learning and Verbal Behavior 7, 291–295.

Larochelle, S., Pineau, H., 1994. Determinants of response times in the semantic verification task. Journal of Memory and Language 33, 796–823.

Larochelle, S., Richard, S., Souli?res, I., 2000. What some effects might not be: The time to verify membership in "well-defined" categories. The Quarterly Journal of Experimental Psychology 53A, 929–961.

Link, S. W., Heath, R. A., 1975. A sequential theory of psychological discrimination. Psychometrika 40, 77–105.

Loftus, E. F., 1973. Category dominance, instance dominance, and categorization time. Journal of Experimental Psychology 97, 70–74.

McCloskey, M. E., 1980. The stimulus familiarity problem in semantic memory research. Journal of Verbal Learning and Verbal Behavior 19, 485–502.

McCloskey, M. E., Glucksberg, S., 1978. Natural categories: Well defined or fuzzy sets? Memory & Cognition 6, 462–472.

McCloskey, M. E., Glucksberg, S., 1979. Decision processes in verifying category membership statements: Implications for models of semantic memory. Cognitive Psychology 11, 1–37.

Morrison, C. M., Ellis, A. W., Quinlan, P., 1992. Age of acquisition, not word frequency affects object naming, not object recognition. Memory & Cognition 20, 705–714.

Morrison, C. M., Gibbons, Z. C., 2006. Lexical determinants of semantic processing speed. Visual Cognition 13, 949–967.

Nosofsky, R. M., Palmeri, T. J., 1997. An exemplar-based random walk model of speeded classification. Psychological Review 104, 266–300.

Ratcliff, R., 1978. A theory of memory retrieval. Psychological Review 85, 59–108.

Ratcliff, R., 1979. Group reaction time distributions and an analysis of distribution statistics. Psychological Bulletin 86, 446–461.

Ratcliff, R., Rouder, J. N., 1998. Modeling response times for two-choice decisions. Psychological Science 9, 347–356.

Ratcliff, R., Schmiedek, F., McKoon, G., 2008. A diffusion model explanation of the worst performance rule for reaction time and IQ. Intelligence 36, 10–17.

Ratcliff, R., Smith, P. L., 2004. A comparison of sequential sampling models for two-choice reaction time. Psychological Review 111, 333–367.

Ratcliff, R., Tuerlinckx, F., 2002. Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. Psychonomic Bulletin & Review 9, 438–481.

Rips, L. J., Shoben, E. J., Smith, E. E., 1973. Semantic distance and the verification of semantic relations. Journal of Verbal Learning and Verbal Behavior 12, 1–20.

Rosch, E., 1973. On the internal structure of perceptual and semantic categories. In: Moore, T. E. (Ed.), Cognitive development and the acquisition of language. Academic Press, New York, pp. 111–144.

Rouder, J. N., Lu, J., 2005. An introduction to Bayesian hierarchical models with an application in the theory of signal detection. Psychonomic Bulletin & Review 12, 573–?604.

Rouder, J. N., Lu, J., Speckman, P. L., Sun, D., Jiang, Y., 2005. A hierarchical model for estimating response time distributions. Psychonomic Bulletin & Review 12, 195–?223.

Rouder, J. N., Lu, J., Sun, D., Speckman, P., Morey, R., Naveh-Benjamin, M., 2007. Signal detection models with random participant and item effects. Psychometrika 72 (4), 621–642.

Rouder, J. N., Speckman, P. L., 2004. An evaluation of the Vincentizing method of forming group-level response time distributions. Psychonomic Bulletin & Review 11, 419–?427.

Ruts, W., De Deyne, S., Ameel, E., Vanpaemel, W., Verbeemen, S., Storms, G., 2004. Dutch norm data for 13 semantic categories and 338 exemplars. Behavior Research Methods, Instruments, & Computers 36, 506–515.

Smith, E. E., Shoben, E. J., Rips, L. J., 1974. Structure and process in semantic memory: A featural model for semantic decisions. Psychological Review 81, 214–241.

Steyvers, M., this issue. Combining feature norms and text data with topic models. Acta Psychologica, this issue.

Thomas, E. A. C., Ross, B. H., 1980. On appropriate procedures for combining probability distributions within the same family. Journal of Mathematical Psychology 21, 136–152.

Tuerlinckx, F., 2004. The efficient computation of the distribution function of the diffusion process. Behavior Research Methods, Instruments, & Computers 36, 702–?716.

Vandekerckhove, J., Tuerlinckx, F., 2007. Fitting the Ratcliff diffusion model to experimental data. Psychonomic Bulletin & Review 14, 1011–1026.

Vandekerckhove, J., Tuerlinckx, F., 2008. Diffusion model analysis with MATLAB: A DMAT primer. Behavior Research Methods 40, 61–72.

Vandekerckhove, J., Tuerlinckx, F., Lee, M. D., 2008. A Bayesian approach to diffusion process models of decision-making. In: Sloutsky, V., Love, B., McRae, K. (Eds.), Proceedings of the 30^{th} Annual Conference of the Cognitive Science Society. Cognitive Science Society., Austin, TX, pp. 1429–1434.

Vandekerckhove, J., Tuerlinckx, F., Lee, M. D., 2009. Hierarchical diffusion models for two-choice response times, manuscript submitted for publication.

Voss, A., Voss, J., 2007. Fast-dm: A free program for efficient diffusion model analysis. Behavior Research Methods 39, 767–782.

Wagenmakers, E.-J., in press. Methodological and empirical developments for the Ratcliff diffusion model of response times and accuracy. European Journal of Cognitive Psychology.

Wilkins, A. J., 1971. Conjoint frequency, category size, and categorization time. Journal of Verbal Learning and Verbal Behavior 10, 382–385.

Zeigenfuse, M. D., Lee, M. D., this issue. Finding the features that represent stimuli. Acta Psychologica, this issue.

A. Fit of the HDM to the data

In order to give an indication of the absolute fit of the HDM to the data at hand, we generated new data from the HDM with the parameters that we obtained. We feel that a model fails to be useful if it fails to capture some sort of regularity that is present in the data. Comparing generated data to the real empirical data might give an indication of unexplained patterns. Since models are by nature idealized representations, some deviation is always expected, but such deviation should not be systematic.

By generating 20,000 new data sets from the model, we obtain the distribution of each data point, as predicted by the model. Figure 1 shows three example data points.

As it would be uneconomical to clutter this appendix with thousands of pages of graphs,[12] we summarized the relevant aspect of each plot. Figure 2 shows the distribution of deviances of the observed data from its a posteriori mean prediction (negative values indicate that the response was slower than predicted by the model). The eight different distributions correspond to the different semantic categories, but they do not differ in any meaningful way. The deviance distribution is left-skewed, indicating that the data contain more positive outliers than negative ones (as is expected in a latency distribution). In general, the model fits the data quite well.

[12]The full set of graphs may be obtained from the first author upon request.

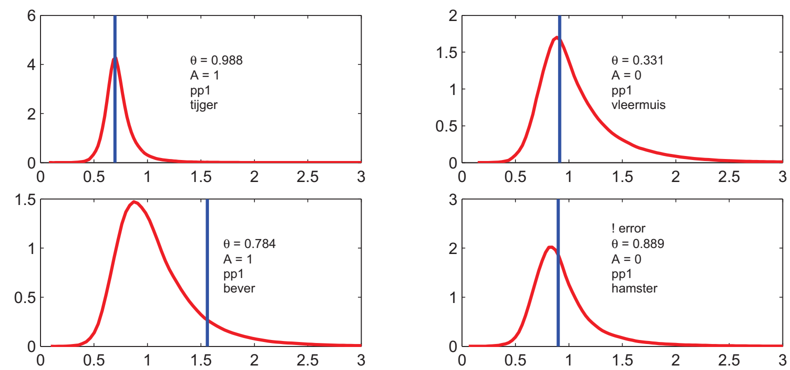

Figure 1: Three example fit graphs for single data points (all taken from participant 1, category ***mammals***). θ is the expected probability of a correct response, and A is the actual response given. The top left graph shows the observed data point (vertical line) exactly on the mode of the posterior predictive distribution, indicating an excellent prediction. On the top right, the model correctly predicts that the participant made an error on the stimulus *bat*. The majority of graphs look like the ones on the top row. The bottom left graph shows a serious latency underestimation by the model. The bottom right graph shows a data point whose response latency is predicted fairly well, but where the model did not predict that participant made a categorization error on *hamster*.

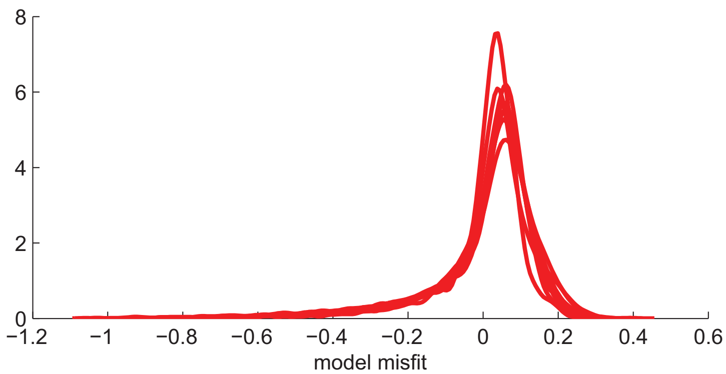

Figure 2: The distribution of model error, per semantic category. See text for details.