Acta Psychologica 144 (2013) 83–96

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/ locate/actpsy

Modeling accuracy as a function of response time with the generalized linear mixed effects model

D.J. Davidson [a,*], A.E. Martin [b]

[a] Basque Center for Cognition, Brain, and Language, Donostia, Basque Country, Spain
[b] University of Edinburgh, Edinburgh, Scotland, United Kingdom

ARTICLE INFO

Article history:
Received 7 December 2012
Received in revised form 11 April 2013
Accepted 24 April 2013
Available online 14 June 2013

PsycINFO codes:
2240
2340
2720

Keywords:
Speed–accuracy tradeoff
Generalized linear mixed effects model

ABSTRACT

In psycholinguistic studies using error rates as a response measure, response times (RT) are most often analyzed independently of the error rate, although it is widely recognized that they are related. In this paper we present a mixed effects logistic regression model for the error rate that uses RT as a trial-level fixed- and random-effect regression input. Production data from a translation–recall experiment are analyzed as an example. Several model comparisons reveal that RT improves the fit of the regression model for the error rate. Two simulation studies then show how the mixed effects regression model can identify individual participants for whom (a) faster responses are more accurate, (b) faster responses are less accurate, or (c) there is no relation between speed and accuracy. These results show that this type of model can serve as a useful adjunct to traditional techniques, allowing psycholinguistic researchers to examine more closely the relationship between RT and accuracy in individual subjects and better account for the variability which may be present, as well as a preliminary step to more advanced RT–accuracy modeling.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

Response time and accuracy are both common dependent measures in experimental psycholinguistics and cognitive psychology. Most often these two variables are analyzed separately, with the implicit (and untested) assumption that they represent two independent response measures, though they issue from the same underlying process or processes. However, the existing literature shows that they are often not statistically independent of each other, and moreover the relationship is sometimes complex, and subject to individual differences. Since RT and accuracy are variables derived from the same decision process with an unknown and dynamic criterion, it seems conceptually difficult to regard them as independent variables, and ideally, statistical models of RT and accuracy should reflect this (Fitts, 1966; Pachella & Fisher, 1969; Pachella & Pew, 1968; Pew, 1969; Ratcliff, 1985; Ratcliff & Hacker, 1981; Ratcliff & Rouder, 1998; Wickelgren, 1977). Many researchers consider qualitatively whether a tradeoff between RT and accuracy is present in their data, at least at the level of group or condition averages. Indeed, this represents an additional "researcher degree of freedom" in the analysis of many data sets (Simmons, Nelson, & Simonsohn, 2011), because researchers can choose whether to emphasize either the results of the RT or the error analysis in support of their claims, when in fact, the two are often not independent sources of evidence. However, even when a tradeoff is not present, an accurate model of the relationship between RT and accuracy may improve the statistical analysis of a given data set. Recent work in psychometrics by Loeys, Rosseel, and Baten (2011), building on earlier work by Van der Linden (2007), has shown how to construct a joint linear mixed effects model for the RT–accuracy relation using a Bayesian approach. The present paper provides a simplified mixed effects model that can be used as a building block for these more elaborate analyses. We argue that classifying individual subjects' relationship between error rate and RT, as well as the group-level pattern, is an important first step in data analysis that can offer critical insights for common psycholinguistic paradigms.

Broadly speaking, there may be one of three (simple) relationships between the time it takes for participants to respond, and the probability that they make an error on a given trial. First, it can be that the more accurate subjects are, the earlier they respond (or, equivalently, with decreasing accuracy they respond later). In this situation lower error rates and earlier RTs both indicate better performance in some sense. This means that participants are not trading response time for accuracy.

A second type of relationship is for subjects to become more accurate at the *expense* of response time. That is, the more accurate subjects are, the slower they respond. This is more commonly known as a speed–accuracy tradeoff, and this pattern is particularly problematic when two or more experimental conditions are to be compared. If participants are more accurate but also later in one condition compared to another, one must entertain the possible explanation that there is not a simple effect of condition on accuracy or RT, but rather a more complex effect of condition on the RT–accuracy relation. This

* Corresponding author. Tel.: +34 943 309 300 209.
E-mail address: d.davidson@bcbl.eu (D.J. Davidson).

0001-6918/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.actpsy.2013.04.016

does not invalidate RT or accuracy as response measures, but depending on the magnitude and direction of the tradeoff, it can be difficult to draw conclusions about a dataset.

Finally, there might be no systematic relationship between RT and accuracy. In this case, a curve relating accuracy to RT will be essentially flat. Although there are important exceptions, this relationship appears to be the most commonly-assumed scenario for psycholinguistics, cognitive psychology and cognitive neuroscience researchers, at least implicitly, because it is currently the most commonly accepted practice for the analysis of RTs and errors to be presented as if the effects are independent of each other. However, it is still relatively rare for researchers to formally test whether this is the case.

In all three of these scenarios, the relationship between accuracy and RT can be defined at either the *subject* level or the *trial* level of analysis. At the level of subject averages, some subjects may be faster as well as more accurate, while others may be faster only when they are less accurate, or there may be no systematic relationship between average RT and average accuracy. At the trial level the relation is defined between the probability of responding correctly on an individual trial and the individual trial RT, rather than the average accuracy versus average RT. These two levels of analysis need not have the same relation. Even if it is the case that (on average) fast subjects are not any more likely to be more accurate, it can be the case that each subject shows a systematic relation between RT and accuracy around their individual subject-level averages. This important distinction between these two levels of analysis is discussed in more detail in Appendix B.

Besides the three possibilities described above, another reason the relationship between RT and accuracy is complex is that it is not always a linear relationship, in the sense of a straight line. When participants are making relatively many errors or relatively few errors, they may still take a short or a long amount of time to respond. That is, large differences in response time may correspond to relatively small changes in proportion correct (and vice versa). The probability of responding correctly or incorrectly is constrained between 0 and 1, but the time taken to respond in a task is typically constrained only by instructions or a response deadline, if at all. The result of this is that often the RT–accuracy relation has the form of a curve, and it is not well modeled using ordinary linear regression, without transforming the variables in some way. This RT–accuracy curve can, however, be modeled with logistic regression as we will outline below.

In sum, the relationship between accuracy and RT, when present, is sometimes not a simple linear function, and there are multiple levels to the relation. Most studies treat accuracy and RT as independent response measures, or arrange the experimental situation so that participants have a relatively high accuracy rate. However, in cases where participants have relatively low or relatively high accuracy, small changes in accuracy can correspond to large differences in response time.

1.1. RT–accuracy tradeoff functions

The notion that people can trade response time for accuracy in any task has been well-documented (Fitts, 1966; Garret, 1922; Hick, 1952; Ollman, 1966; Pachella & Pew, 1968; Pew, 1969; Schouten & Bekker, 1967; Wickelgren, 1977; Woodworth, 1899). At the heart of the problem is the fact that individual participants respond per an unknown internal criterion that is likely to be dynamic over time. Thus, participants can trade the speed of response for accuracy of response based on unobservable changes or differences in internal criterion. In order to study the timecourse of information processing, it is therefore more informative to obtain a full RT–accuracy function for an individual performing a given task, of which an RT would yield only one point in time. Wickelgren (1977) outlines various experimental procedures to derive the function (payoffs, deadlines, instructions, response binning or partitioning, and lastly the application of response signals) and argues that the only way to prevent speed–accuracy tradeoff is to use the specialized response-signal interruption paradigm (Reed, 1973, 1976; Schouten & Bekker, 1967). Unfortunately, the specialized design and the procedure needed to implement such a paradigm are not always feasible, nor desirable to many researchers. Furthermore, the analysis strategies are specialized — requiring special designs or statistical estimation techniques. For example, there are limitations as to the interpretation of partitioned responses, and often problems with sparse data in early bins of short reaction times — see Wickelgren (1975), Wickelgren (1977) and Schouten and Bekker (1967). Here, we advocate a simpler approach to diagnosing whether there is a tradeoff, or not, between RT and accuracy in a given dataset, without the application of specialized designs or procedures. Our aim is to give the user a straightforward and simple method for assessing the statistical relationship between RT and accuracy in a dataset — our approach is agnostic regarding the model of the underlying decision process that leads to performance and to the particular relationship between the two variables. However, we note that our analysis approach shares the same core assumptions seen in the extensive literature on computational and theoretical models of two-alternative forced-choice decision processes (e.g., Ratcliff, 1978; Ratcliff, Gomez, & McKoon, 2004; Ratcliff & McKoon, 2008) — namely that response time and performance are inextricably linked, that the relationship between the two must be included in any statistical model of the data, and that the former and the latter points are crucial for interpretation of the data. Note that the existence of any computational and/or statistical relationship (in our case, only the latter) between RT and accuracy in no way implies that changes in RT cause changes in accuracy, or vice versa, if RT is modeled as a function of accuracy.

An important observation to make is that both response time and accuracy can be modeled as random effects in the sense that a typical sample of response times or response choices will have a statistical *distribution*. This distribution can depend strongly on the particular subject who has been sampled. If this is the case, then at the level of individual trials of an experiment, there should be a strong relationship between the response time and the response choice because the same subject variability is affecting both, but is independent of other subjects. That is, RTs should be informative about accuracy at the trial level because both dependent measures will reflect individual variation in participants. At the same time, there may also be a systematic relation between RT and accuracy at the group level. The next section describes our approach to modeling RT as a regression input at these multiple levels.

1.2. Linear and generalized linear mixed effects models

In this paper we will model the proportion response as a function of RT, where the RT enters the model as either a fixed and/or a random effect. This is analogous to the situation in many datasets where the performance outcome variable y is binary. Examples include correct/error response, present/absent decisions about a stimulus, recalled/not-recalled in a memory experiment, or fluent/disfluent in a production experiment. All of these examples share the essential characteristic that the response y takes on one of two values. This situation is different from a response measure like RT, because the binary response is not accurately modeled as a Gaussian distribution at the trial level (e.g., in cases where the response is actually distributed as a binomial). In logistic regression (see Jaeger, 2008 for an introduction, also Jaeger, Graff, Croft, & Pontillo, 2011; Quené & van den Bergh, 2008), we instead model the probability that response $= 1$ for some regression input x in terms of the inverse logit: $Pr(y = 1|x) = \exp(\beta x)/(1 + \exp(\beta x))$ where x is parameterized with the coefficients β to represent the effect of experimental variables, as well as variables like RT. Here, we use the inverse logit because usually one wants to go from a calculated coefficient in our model to proportions,

and this is easier to define in terms of the inverse logit. Rather than a straight-line relationship between the regression input and the outcome, with the inverse logit we have a curve. The curvature towards 0 and 1 with increasing (decreasing) values of x is necessary to keep the probability estimate bounded between 0 and 1, and in our case, to model the RT–accuracy curve. Logistic regression can be performed using generalized linear regression models (GLMs), or as in the present paper, mixed effects GLMs.

Mixed effects linear models are one type of model appropriate for data from repeated measures experiments because the data from these experiments is *grouped* by subject (and/or other factors like item). See Baayen, Davidson, and Bates (2008), as well as Quené and van den Bergh (2004, 2008) for introductions, and also Kliegl, Masson, and Richter (2010a) and Kliegl, Wei, Dambacher, Yan, and Zhou (2010b) for some recent applications. Data from the same subject will tend to be correlated because there are subject characteristics (e.g., response criterion, response bias) which differ between subjects but are the same for a given subject. Ordinary linear models do not model these grouping characteristics and as a result underestimate the variation in the data. The mixed effects logistic regression is an extension of the linear mixed effects approach for logistic regression (see Bates, 2010 and Jaeger, 2008 for an introduction).

Mixed effects GLMs can be used to model multiple grouping variables, and in addition, correlations present in the random effect variation. In the present paper, this is used to model subject-specific deviations from the fixed effect estimates of the response, with respect to RT. That is, RT can be modeled *both* as a fixed effect (a relationship estimated over all participants), and as a subject-specific deviation from the fixed effect estimate. For example, if the response variable is a measure of whether a participant correctly recalled an item on a given trial, the trial-specific RT can be used to account for this response both as a fixed effect (a general effect of RT, modeled as the slope of a regression of RT on the probability of recall) and a random effect (a subject-specific deviation in the slope of the recall–RT function).

A recent approach to this problem by Loeys et al. (2011) uses a *joint* modeling approach which models both RT and accuracy with shared covariance between the random effects for RT or accuracy for either subjects or items. In this approach, there are correlation parameters that link the covariance of the subject- or item-random effects, and these correlations can be used to assess the relationship between RT and accuracy. Although it was not presented in Loeys et al. (2011), it would be relatively straightforward to extend their approach to assess interactions between random effect deviations from a condition effect in the correlation between RT and accuracy, to address some of the analysis goals outlined in the present paper. However, as the authors themselves emphasize, their approach is relatively labor intensive with respect to modeling, as it requires setting up a Bayesian model for the joint relation and simulation modeling of the response. While there are many reasons to prefer this approach, it can also be useful to adopt a simpler mixed effects approach in the initial stages of an analysis, before constructing a full Bayesian model (see, e.g., Gelman & Hill, 2007, p. 345). Second, as we will show below, the approach of Loeys et al. (2011) is a model of random effect variation over subjects (or items). This corresponds, roughly speaking, to a relationship between mean RT and mean accuracy, calculated over subjects (or items). However, as is well known, it is also possible that accuracy and RT are related over trials within an individual subject (or item). The classical literature on speed–accuracy tradeoff, for example, has emphasized the accuracy–RT relation within individual subjects, for example. The goal of this paper, therefore, is to present a simplified approach to modeling accuracy based on using RT as *both* a fixed and a random effect regression input that will allow researchers to characterize the accuracy–RT relation at both subject (or item) and trial levels. Note that one could also model RT as a function of accuracy (e.g., Rabbitt, 1967) — here we concentrate on accuracy as a function of RT to simplify the presentation, and because (informally) it appears that the generalized linear (mixed) model is not yet widely considered for single-trial accuracy data. Our approach has the added benefit that it may also appeal to researchers in a given field who do not wish to make detailed cognitive assumptions in their choice of procedure or analysis technique (see earlier discussion of the SAT approach). This would apply particularly for research topics where detailed knowledge about cognitive states or processes is not available, but nevertheless a technique that can assess how RT and accuracy are related to each other is required.

2. Example data analysis 1

The example data set that we use to illustrate the technique advocated in this paper is a memory recall task using spoken words. Spanish-native (L1) speakers learned to translate Spanish nouns into Basque (L2) in a translation learning task in which both accuracy and response time were measured. The central result was an effect of translation direction such that memory performance was greater with L2-cued responses, compared to L1-cued responses. That is, subjects could translate more accurately when they translated from the newly-learned Basque into their native Spanish, compared to translation from Spanish into Basque. This effect has been demonstrated previously (e.g., Kroll & Stewart, 1994), but in most cases the RT and accuracy have been analyzed independently of each other, following the usual conventions for analyzing this type of data. The purpose of the analysis presented here is to investigate at the trial level how each participant's response time was related to the rate of recall in order to illustrate how our proposed analysis would work with an actual data set. Simulations of other scenarios with simple RT–accuracy relations are presented in the sections following the example data set.

2.1. Method

2.1.1. Participants

Twenty-two native speakers of Spanish with no or very little (self-reported) knowledge of Basque took part in the experiment. All participants reported learning Spanish from their parents and reported that they did not know Basque, and that they had not taken courses to learn Basque. All participants were right-handed and had no previous history of hearing or neurophysiological impairment. Participants were given monetary compensation for the experiment. As part of the informed consent procedure, all participants were informed about the instructions and task in the experiment and signed a consent form before taking part. The experiment was conducted in Spanish and all of the task instructions and administrative materials were in Spanish.

2.1.2. Materials and design

The materials consisted of recordings of spoken Spanish nouns and spoken Basque nouns. The recordings were made using two fluent female Basque–Spanish bilingual speakers in their mid-20s, using a high quality microphone directly digitized to disk at a sampling rate of 44 kHz via software ('Praat', see Boersma & Weenink, 2005).

The spoken word pairs were arranged in six pairs per list. Participants first heard each pair in an encoding phase, and then were provided with one member of the pair (either Basque, or Spanish) as a probe in a retrieval phase. The encoding and retrieval practice repeated three more times each list, so that participants had four attempts at recall total for each pair. In all there were eight lists for a total of 48 pairs.

The nouns were chosen so that the length and phonological complexity of each pair would be comparable. Several databases were used to find the word pairs. The 'apertium' translation database

(Tyers, Sánchez-Martínez, Ortiz-Rojas, & Forcada, 2010) was used to define an initial set of Basque–Spanish noun translations. Next, using the citation-form phonological transcription in lexical databases for Spanish ('B-Pal', Davis & Perea, 2005) and Basque ('E-Hitz', Perea et al., 2006), pairs of nouns were chosen such that: the members of each pair had the same number of syllables but distinct initial phonemes, the Levenshtein distance between the CV transcriptions of each pair was less than three, the Levenshtein distance between the phonological transcriptions was greater than three (to avoid cognates), the absolute difference in $\log_{10}$ frequency was less than 2, both the Spanish and Basque frequencies-per-million were greater than 5, both the Spanish and the Basque words had a noun part-of-speech tag, and that the absolute difference in the number of phonemes was not greater than 1. Note that the lexical databases for Spanish and Basque are compiled from written sources. From this set of 250 pairs, the candidate words were reviewed and selected by a Basque–Spanish bilingual to be included in the final experimental lists.

During the stimulus recordings, the Spanish nouns were produced with a preceding definite article (e.g., 'la casa', the house), appropriate for the noun, in order to be comparable to the Basque nouns, which require a following article (e.g., 'etxea', the house). Note that Spanish has nominal gender agreement, so that the article agrees in gender with the noun, while Basque does not have nominal gender agreement. Basque nouns ending in '-a' (so-called organic '-a') were not used in the stimulus lists so that the length of the Spanish and Basque terms would be comparable after addition of the article.

2.1.3. Procedure

The experiment was carried out as part of an MEG (magnetoencephalography) recording, and for this reason a cued-response task was used. Muscle artifact generated during speech can disrupt MEG recordings, and therefore subjects were asked to provide their translations in the retrieval phase in response to a visual and auditory cue (a question mark '?' and a brief tone) presented 1 s after the offset of the retrieval probe.

Participants first heard the six pairs of words in an encoding phase in which each word was presented successively (1 s ISI between the offset of one word and the start of the next within a pair; 3 s ISI between the offset of one word pair, and the onset of the next pair). Then there was a brief interval in which participants were asked to blink their eyes 10 times in a row (this was done to reduce artifact in the MEG recordings, and to provide a counting task between encoding and retrieval). After the blink period, a retrieval phase started in which participants heard the retrieval cues and provided their (spoken) responses. They first heard a brief (0.1 s) warning tone and saw a fixation cross ('+') for 1 s, followed by the presentation of a recall cue word. For 1 s after the onset of the recall cue, the fixation remained on screen, and then was replaced by a question mark ('?') which remained on screen for 5 s or until the experimenter coded the response. When the experimenter coded the response, a green square flashed on the screen for 0.25 s to indicate a correct response, or a red square (also for 0.25 s) for an incorrect response. The next retrieval cue trial followed immediately after the feedback. There were six retrieval trials, corresponding to the word pairs heard during encoding. The retrieval trials alternated between either Spanish or Basque within each retrieval test. That is, for a given set of six retrieval probes in a retrieval test, all of the probes were Spanish or they were all Basque. The retrieval trials were arranged so that two of the tests were in Spanish, and two were in Basque for each list. The position of the Spanish- or Basque-cue for retrieval in a given list was assigned to each participant via a Latin square so that the Spanish and Basque cued trials occurred an equal number of times across lists and participants at each of the four retrieval practice positions per list. The order of the word pairs during encoding, and the order of the retrieval cues during retrieval was randomized individually for each participant with the constraint that words that ended the encoding phase would not be immediately tested in the retrieval phase (to avoid a recency effect). There were different random orders for each encoding and retrieval phase.

There was no special emphasis to provide a speeded response in the task. Subjects were told to try to produce the correct translation after the response cue appeared, or to produce 'supongo' (don't know) if they could not recall the translation. Participants were asked to produce the full form of the nouns, including the articles, in the same form that they heard them during encoding. The response interval was 5 s, and the response time was measured from the presentation of the cue.

2.2. Results

Fig. 1 shows the average proportion correct recall for each participant as a function of (quantiles of) log response time to produce the translations. This corresponds (approximately) to the trial-level relationship between recall and RT in each participant. In nearly all participants, recall performance was higher for the Basque-cued trials compared to the Spanish-cued trials. Also, in nearly all participants (in both conditions), recall performance was higher when response times were faster. That is, the trend of response accuracy with increasing log RT is generally negative. Below, several models of this relation are evaluated.

First, however, we show the relationship at the level of subjects. Fig. 2 shows the average proportion recall as a function of log response time over subjects, for the two translation directions. This differs from Fig. 1 because each participant contributes only two points corresponding to their average level of performance (one for Basque-cued, and another for Spanish-cued trials). At this level of analysis, in contrast to Fig. 1, there is a relatively weak relationship between recall accuracy and RT, if any. For the data shown in Fig. 2, for the Basque-cued trials the Pearson product–moment correlation was $r = -0.190$, and the 95% CI -0.566, 0.252 included zero. For the Spanish-cued trials, the correlation was slightly higher, $r = -0.366$, but the 95% CI -0.682, 0.066 also included zero. There was also no evidence of a difference between these correlations ($t = -0.17$). Thus, there was no strong evidence of a relationship between recall accuracy and RT at the level of subject averages.

At the level of trials within a subject, Fig. 1 suggests that response time is statistically related to recall performance. However, the proportions shown in Fig. 1 are aggregated over trials that are grouped into RT-quantiles. To estimate the effects of translation direction and RT using *single* trial data, a mixed effects logistic regression model of the response was constructed for these factors using participant and item as random effects. The models were fit using a Laplacian approximation to the log-likelihood. Four different models were compared, all with the same fixed effects specification (simple effects for translation direction and RT, and the interaction of translation direction and RT, using "treatment coding" in R). Table 1 shows the different random effects specifications. The models only differed with respect to the random effects formulation for log response time.

The first model **m0** included simple random effects for item and subject, as well as a translation–direction random effect for subjects and a correlation parameter for the intercept and the condition random effects. There was no random effects model term for log RT. This model was intended to capture individual variation in the effect of translation direction, as well as any correlation between the (subject-specific deviation in the) overall level of recall and the size of (subject-specific deviation in) the translation–direction effect.

The second model **m1** included all of the random effects of model **m0**, and in addition, included a subject random effect slope for (centered) log RT, uncorrelated with the other random effects in the model. This model was intended to capture individual variation in the recall–RT relation beyond the population-level effect of RT on

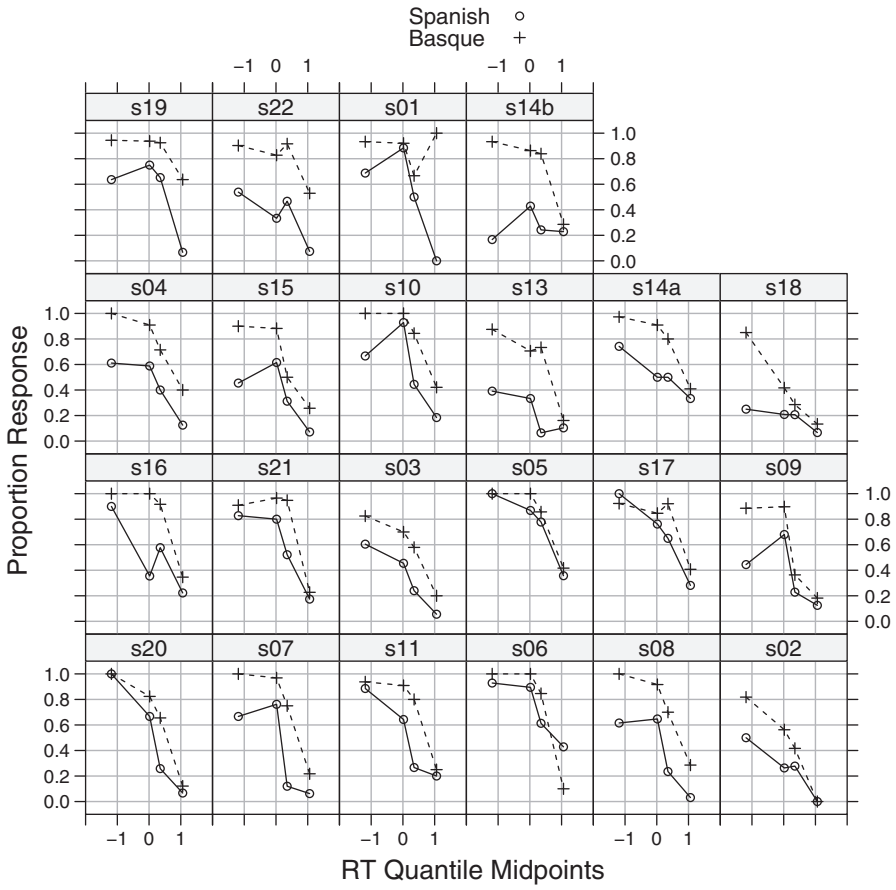

Fig. 1. Proportion recall as a function of (log) response time for each participant in the translation–recall experiment. Dotted lines (+) correspond to translation from Basque to Spanish, and solid lines (○) correspond to translation from Spanish to Basque. The proportion recall for each participant was calculated at quantiles of the (log) RT over participants.

recall, but it does not model whether the individual variability in the RT–recall relation is related to the individual variability of the translation–direction effect.

The third model **m2** was like model **m1**, but in addition included a correlation parameter between the (centered) log RT (slope) random effect and the other subject random effects in the model (Pinheiro & Bates, 2000; and Section 3.2 of Bates, 2010). This model is intended to model relationships like **m1**, but in addition capture correlations in the subject-level variability. Here, the different random effects in the model, like the subject-specific deviation from the group-level estimate of proportion correct, or the subject-specific variation in the slope of the RT–recall function, are modeled with a correlation coefficient between the random effects. Correlation coefficients like this could, for example, diagnose whether subjects who have higher level of performance also have a stronger relation between RT and recall.

The final model **m3** was like model **m2**, but also included an interaction term between translation direction and (centered) log RT in the subject random effects specification. This model was intended to capture individual differences in the slope of the RT–accuracy function that are modulated by the direction of recall (i.e., condition-specific recall–RT slopes in each subject).

Table 2 shows the fixed effect parameter estimates (logit probability) for the four models. All of the fixed effect contrasts in the models were statistically significant (all z values for greater than 5), and as Table 2 shows, the four models gave similar effect estimates: there were main effects of both translation direction and log RT, as well as an interaction between the two, consistent with Fig. 1. Taking the inverse logit for model **m2**, for example, at the mid-point of log RT (log 1 s), there was almost 40% improvement in recall when translating in the easier condition (L2 to L1; 0.900) compared to the more difficult condition (L1 to L2; 0.504). However, this effect was strongly related to RT: for a one-unit increase (approximately 2.7 s) from the mid-point in log RT, recall declined substantially in both the easier condition (L2 to L1; 0.552) and even more so in the harder condition (L2 to L1; 0.254). Thus, subjects found it easier to translate from Basque to Spanish, and they were more successful when doing so if their answers were relatively fast compared to when their answers were slow. This is not consistent with speed–accuracy tradeoff, for example, because the parameter estimates indicate that participants were *less* accurate when they were slower.

A comparison of the four different models also indicated that there was substantial individual variation in the relationship between RT and the response, as suggested by Fig. 1. Table 3 shows the sequential model comparison statistics comparing the three models that included log RT in the random effects specification as well as the null model (**m0**) without log RT as a random effect. The comparison among these models indicates strong support for the choice of either model **m1** or **m2** over the null model, indicating that the slope of the recall–RT function differs between subjects.

Model **m1** included a simple additive slope of RT, while model **m2** also included correlations between the random effect deviations. Because the fixed effect parameter estimates are largely similar for these two models, it seems more parsimonious to choose model **m1** as a representation of the data. The value of model **m2** is showing that the effect estimates nevertheless remain similar to the simpler model when modeling the correlations between individual differences.

The last model (**m3**) does not show a better fit than either **m1** or **m2**, suggesting that there was not a great deal of individual *variation*

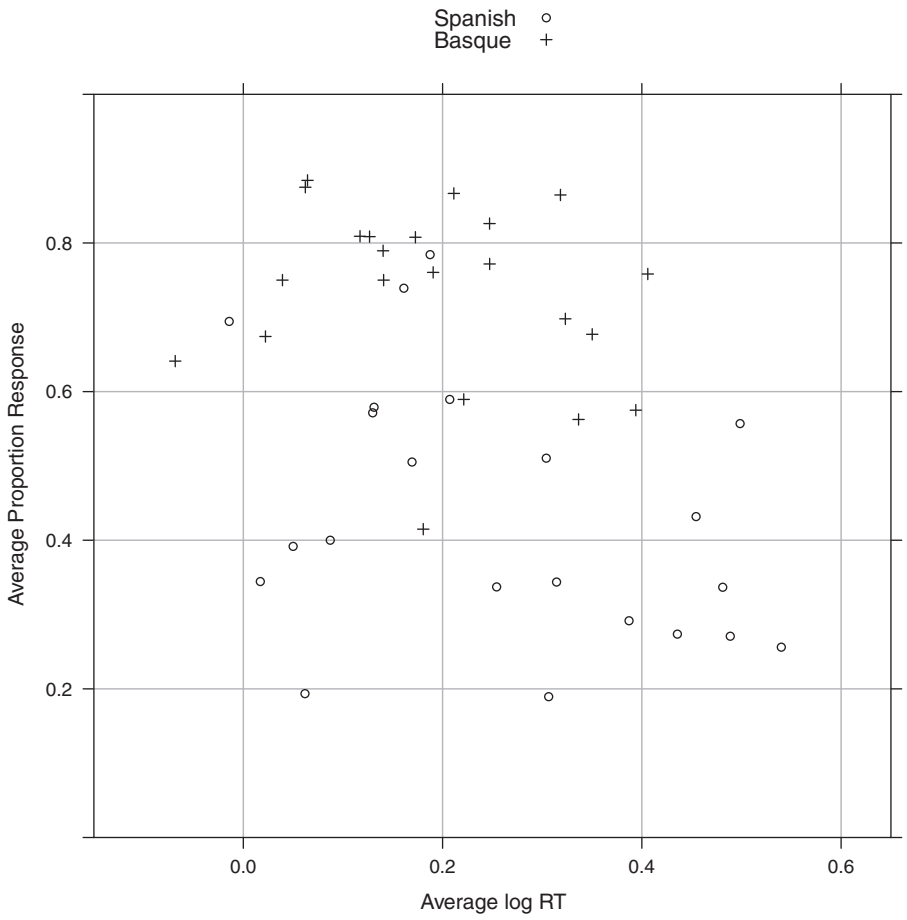

Fig. 2. Average proportion recall as a function of average (log) response time for each participant in the translation–recall experiment. Crosses correspond to translation from Basque to Spanish, and circles correspond to translation from Spanish to Basque.

in the recall–RT slope depending on translation direction. Recall that the fixed effect parameter estimates showed an interaction for participants as a whole. The comparatively small improvement in fit for the last model indicates that there was not a great deal of subject-specific variation for this interaction.

2.3. Discussion

The results showed that translation from L2 to L1 was more successful than translation from L1 to L2 in these early adult Spanish learners of Basque. Second, there was a substantial (negative) relationship between the recall response time and the probability of recall: the faster subjects responded, the more likely they produced the correct translation. Finally, models of the data that included individual variation of this recall–RT relationship better fit the data than a model that did not include the individual variation. Also, the individual variation in the modulation of the recall–RT curve by translation direction was not substantial.

The general effect of translation direction is consistent with previous studies of already-established bilingual participants (Kroll & Stewart, 1994). The results suggest a similar asymmetry in Spanish learners of Basque who are just beginning to learn Basque. The analysis presented here also shows that this asymmetry in recall rate holds when (statistically) controlling for the effects of response time in each subject. In previous studies of this effect, response time and accuracy have been treated as independent variables in the analysis, as is commonly done for this type of data.

With respect to the analysis technique we are advocating in this paper, the analyses presented here showed that response time was negatively related to the rate of recall for participants as a whole. This negative relationship could be observed in most all individual participants (e.g., see Fig. 1). The statistical models of the recall showed that the parameter estimates for the effect of translation direction remained largely similar when controlling for individual variation in the recall–RT relation at the trial level. The combination of the graphical display of the recall–RT relation along with the statistical model(s) for the relationship was effective in ruling out the possibility that the recall asymmetry in translation direction could be explained by either group-level or individual-level variance in recall response times. The graphical display indicated the approximate trend of the recall–RT relation in individual subjects, and the statistical models provided estimates of the slope of the recall–RT function, as well as a statistical assessment of the individual variance of this relationship.

Table 1
Model formulas for the four different random effects specifications. Translation direction (dir) and log RT (rt) are used as regression inputs for the response (rsp), using subject (sbj) and item (itm) labels as random effects.

Model	Formula
m0	`rsp ~ dir*rt + (1 + dir|sbj) + (1|itm)`
m1	`rsp ~ dir*rt + (1 + dir|sbj) + (0 + rt|sbj) + (1|itm)`
m2	`rsp ~ dir*rt + (1 + dir + rt|sbj) + (1|itm)`
m3	`rsp ~ dir*rt + (1 + dir*rt|sbj) + (1|itm)`

Table 2
Fixed effect parameter estimates (standard errors in parentheses) for the four models for the example data set. β_0 corresponds to the logit probability of correctly translating from L1 to L2 (at the mid-point of log RT). β_1 is the simple effect of translating from L2 to L1 (i.e., the difference from β_0). β_2 is the relationship of recall probability with log RT for the translation from L1 to L2. The interaction term $\beta_1\beta_2$ models the change in the slope (from β_2) of the recall–RT function when translating from L2 to L1.

Model	β_0	β_1	β_2	$\beta_1\beta_2$
m0	0.0862 (0.2180)	2.1814 (0.1358)	−1.8260 (0.1270)	−1.3344 (0.2072)
m1	0.1409 (0.2264)	2.1768 (0.1326)	−1.9824 (0.1976)	−1.2702 (0.2063)
m2	0.1565 (0.2336)	2.1862 (0.1323)	−1.9919 (0.2074)	−1.2895 (0.2065)
m3	0.1519 (0.2274)	2.2209 (0.1217)	−2.0094 (0.2099)	−1.3488 (0.2539)

There is one caveat with respect to the languages used in this experiment that should be emphasized. An additional difference between the L1 → L2 and the L2 → L1 translation directions is that the Spanish nouns required retrieval of a gender-appropriate determiner, to be produced before the (Spanish) noun. The Basque nouns required a determiner to be produced after the (Basque) noun, but no gender agreement. Thus the L2 → L1 translation direction required both formulating agreement and the initial production of the determiner, while the L1 → L2 translation direction required only the production of the article following the noun. Thus, it cannot be excluded that the difference in recall rates between the translation directions is due to this difference in requirement for the agreement and order of the article, and not the retrieval difficulty for the phonological form that participants experienced with the newly-learned Basque term, or that both factors play a role. The L1 → L2 translation direction is predicted to be more challenging for participants by the retrieval difficulty account, but the agreement and order account predicts that the L2 → L1 translation direction should be more difficult. Note, however, for the purposes of the present paper, it is not necessary to decide between the retrieval-difficulty and article-order and agreement explanations. The purpose of the analysis presented here was mainly to show how response time is related to retrieval success. To illustrate how this modeling approach might reveal the underlying statistical relationship (if any) between RT and accuracy on the individual and group levels, we turn to several simulated data sets, as well as a comparison of the joint modeling strategy described by Loeys et al. and the present approach.

3. Simulation 1

In the data analysis presented in Section 2, including RT as a regression input did not change the fundamental form of the translation direction effect. Individual subjects showed a negative relationship between accuracy and RT. In the case of an empirical data set as above, the parameters are estimated from the data to infer features of the underlying distributions, but the true model remains unknown. Another useful approach to understanding the behavior of the analysis technique that we advocate is simulation (see Appendix 1). The value of the simulation is in demonstrating how the logistic regression identifies patterns that are previously set up. In this section, we simulate data to model a more difficult case in which there is a tradeoff between accuracy and speed. The simulation here will try to show whether this method will identify when a tradeoff is present in the data of individual subjects.

The RT data for each subject, *RT*, is simulated from $N(0,1)$, a normal distribution with a mean of zero and standard deviation of 1. This represents a range of (standardized and log-transformed) RTs. A subject-specific intercept term ($b0$) for the accuracy–speed curve is drawn from $N(2,1)$. This represents the subject-specific accuracy at when the (standardized) average RT is 0, or the midpoint of the accuracy-speed curve. To model a negative relationship between the distribution of the RTs and the accuracy, a subject-specific slope for the speed–accuracy curve, $b1$, is drawn from $N(-b0,1)$. Note that the mean of this distribution is the intercept term multiplied by −1. Thus, when $b0$ is positive, there will be a tradeoff relationship between accuracy and RT (the more accurate subjects are, the slower they are). When $b0$ is negative, a positive relationship between accuracy and RT will be present. The responses, a vector including 0 s and 1 s, are generated using the inverse logit function with $b0 + b1 * RT$

Three generalized linear mixed models are fit to the data. The first (**m0**) includes a population-level regression input *RT* for the accuracy (*Rsp*), but no individual *RT* parameter for each participant. Models **m1** and **m2** include individual parameters for each participant. The first model, **m1** includes independent parameters for intercept and slope, while model **m2** includes correlated parameters for intercept and slope.

As Table 5 shows, the fixed-effect parameter estimates for the intercept and slope, and the correlation between the fixed-effect parameter estimates ($r = -0.659$), of the best-fitting model are closer to the values specified for the simulation than either the model with no RT-slope random-effect ($r = -0.177$), or the model with independent slope and intercept terms for each subject ($r = -0.091$), see also Table 4. Note in particular the (population-level) correlation between the fixed effect slope and average RT is estimated to be $r = -0659$ for the best fitting model, but was close to zero in the other two models. This is a better account of the simulated data because the simulation was set up precisely to induce a negative correlation at the population level.

Fig. 3 below shows the fit of the best-fitting model for each simulated subject. For each simulated participant, the RTs were binned into four quantiles and the average proportion correct was calculated. These averages are plotted as small circles, superimposed on the population-level estimate (blue line, the same in the plot of each subject), and the estimated subject-level accuracy–speed curve from the best-fitting model (red-line, different for each participant). The green reference line is the average accuracy of each simulated participant at RT = 0. The plot shows, as was specified in the simulation, most of the speed–accuracy curves are negative, such that slower RTs are associated with lower proportion correct responses (most of the lines have a negative slope, with one of the simulated subjects (s15) having a positive relationship). Moreover, participants with a higher level of accuracy tend to have a stronger (negative) relationship with speed.

This simulation has shown that the generalized linear mixed effects model can capture negative and positive relationships between accuracy and speed, at both population and individual subject levels. The model can estimate this relation in each participant to identify participants who present a different pattern from the population-level average.

Table 3
Model comparison of four models for the example dataset. The AIC is smallest for model **m2**, and the BIC is smallest for **m1**.

	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
m0	8	3684.13	3734.48	−1834.06			
m1	9	3667.29	3723.94	−1824.65	18.83	1	0.0000
m2	11	3663.64	3732.88	−1820.82	7.65	2	0.0218
m3	15	3667.76	3762.17	−1818.88	3.88	4	0.4225

Table 4
Model comparison of the three models for the first simulation. The deviance scores (AIC) is smaller for model **m2**.

	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
m0	3	1336.11	1352.25	−665.06			
m1	4	1130.33	1151.84	−561.16	207.78	1	0.0000
m2	5	1125.78	1152.67	−557.89	6.55	1	0.0105

Table 5
Parameter estimates for the three models in the first simulation. The values in parentheses indicate standard errors.

Model	β_0	β_1
m0	1.7131 (0.2488)	−1.3531 (0.0887)
m1	2.2960 (0.3293)	−2.0808 (0.4074)
m2	2.3914 (0.3650)	−2.1943 (0.4419)

4. Simulation 2

The first simulation attempted to capture the general relation between speed and accuracy for a situation where there was no comparison between two conditions. The next simulation (see Appendix C) adds a comparison between conditions.

The simulation is set up similar to the first, except that there are additional variables, $x2$ and $b2$. The variable $x2$ is the condition identifier for each subject, taking on values 0 and 1 according to a binomial distribution with mean of 0.5. The variable $b2$ models the subject-specific condition effect. In this simulation, we have set up $b2$ to be identified with $b1$, so that in the model, there is no condition "effect" by itself except for random deviations drawn from a distribution $N(-b1,1)$. This distribution has a mean that takes on the value of the RT–accuracy slope in each subject. That is, $b2$ is a random variable that is directly correlated with the subjects' RT–accuracy tradeoff.

Fig. 4 below shows the fit of the best-fitting model for each simulated subject. As in the first simulation, for each simulated participant, the RTs were binned into four quantiles and the average proportion correct was calculated. These averages are plotted as either small circles or crosses for two different conditions, superimposed on the population-level estimate (thin lines, the same in the plot of each subject), and the estimated subject-level accuracy–RT curve from the best-fitting model (thick lines, different for each participant). The plot shows that most of the RT–accuracy curves are negative, such that slower RTs are associated with lower proportion correct responses. In some simulated subjects however, there is relatively little relationship between RT and accuracy. There is also a condition difference which varies in magnitude between subjects. Please note that we have also conducted a similar simulation including both item and subject random effects in the model, similar to the approach of Loeys et al. (2011), but the conclusions are not substantially different than the present simulation so we omit it here to conserve space.

In the simplest model (**m0**), with a subject random effect but no random slope for RT, the main effects of Condition and RT are estimated to be approximately −2, as specified in the simulation. The fixed effect correlation between Condition and RT parameters was estimated to be $r = 0.120$. For the model with independent random effects (**m1**), the fixed effect correlation was estimated to be $r = 0.060$. However, the more complex model (**m2**) which accounted for the correlation of Condition and RT ($r = 0.791$) proved to be a better model of the data than the models that do not have the correlation (Table 7).

For models 1 and 2 (**m1**, **m2**), the effect of Condition and RT is of a similar magnitude, and near the value of −2 specified in the simulation (Table 6).

The model comparison shows that the model with the correlated random effects of Condition and RT has lower deviance than the other models: either the baseline model with no random effect for RT (**m0**) or the model with independent random effects for Condition and RT (**m1**).

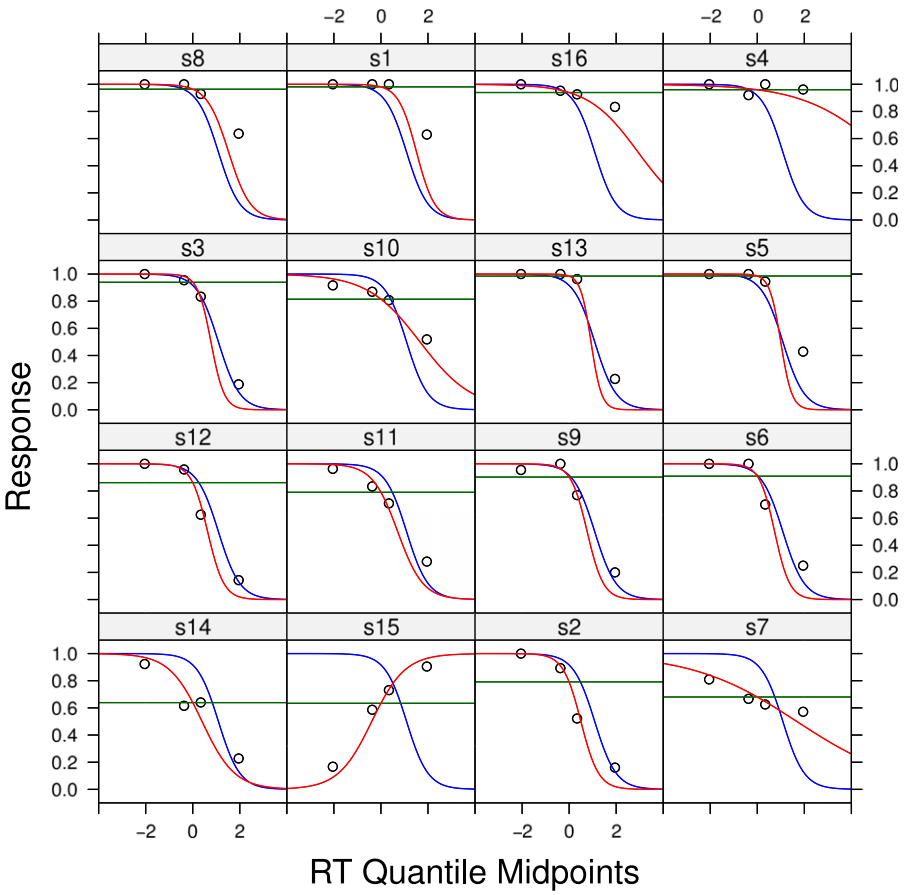

Fig. 3. Data ($N_{Subject} = 16, N_{Trial\ =\ 100}$) from the first simulation. For each (simulated) participant, the horizontal green line indicates the intercept (the response at $RT = 0$), the blue line is the estimated fixed effect (logistic) regression line for the response on the standardized RT, and the red line is the estimated deviation from the fixed effect response. Note that the fixed effect estimate is the same for each participant. The simulated participants are ordered by the intercept estimated by a simple glm.

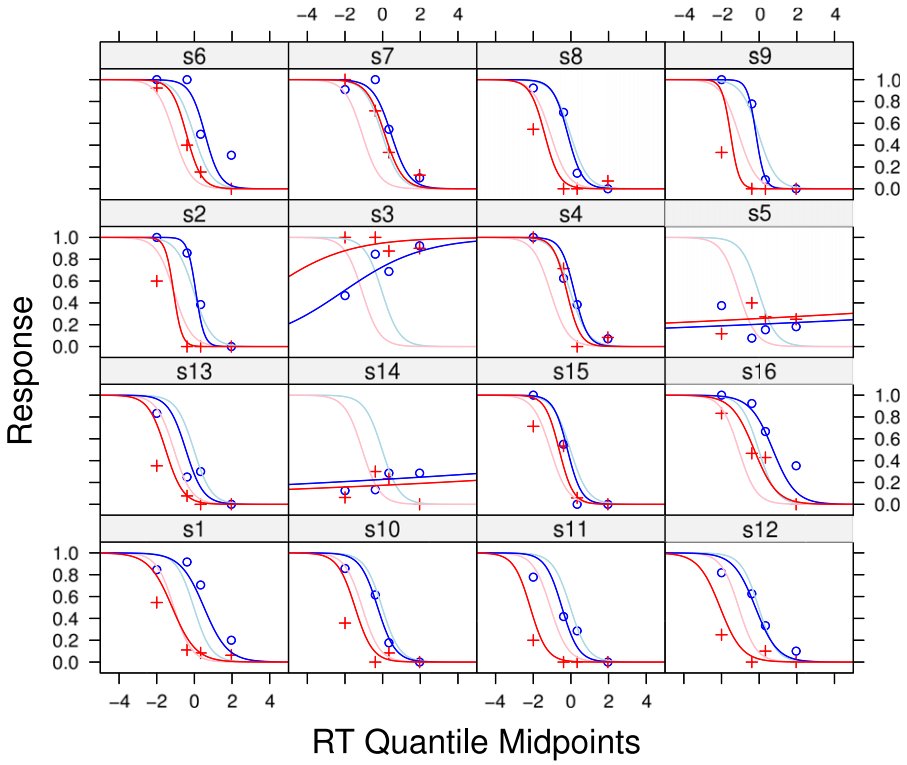

Fig. 4. Data ($N_{Subject} = 16$,$N_{Trial} = 100$) from the second simulation. See the main text for further description.

5. Discussion and conclusions

In this paper, we have modeled the proportion correct response as a function of fixed- and random-effect RT in an attempt to improve the estimate of the response probability by accounting for trial-to-trial covariation between the probability of a response and the RT. The example data set showed that including RT as a regression input resulted in a model that better fits the data. In addition, models that incorporated random-effect variability in the response–RT function better captured the data than those that did not. Two simulations were then presented to show specific cases of how the mixed effects GLM framework can be used to identify the general pattern of the response–RT function over subjects (simulation 1) and the effect of an experimental manipulation on the response–RT function (simulation 2).

Existing work within psychometrics has shown how a more general approach to *jointly* modeling RT and accuracy may be effective in modeling the random effects relation between RT and accuracy (Loeys et al., 2011; Van der Linden, 2007). The goal of the present paper is to make it easier to adopt this approach, by presenting a simplified (generalized) linear mixed effects model for the RT–accuracy relation as a building block for more elaborate models. For the relationship between the two approaches, please see the comparisons in Table 3 of Loeys et al. (2011). As Gelman and Hill (2007) suggest, it can be useful to first construct simplified models as a preliminary to fully Bayesian approaches to linear mixed effects problems. The reason for this is that currently Bayesian models, despite their numerous advantages, require more set-up and evaluation than non-Bayesian approaches. Similarly, if concerns are raised about potential RT–accuracy trade-offs in a given data set, the approach advocated here may serve to quickly identify whether evidence of such a relationship exists, and if so, determine its form. This can be supplemented by a graphical approach of binning responses by quantiles of RT in each participant, as described for example by Wickelgren (1977). These analyses can be used to decide whether a more complex modeling effort is warranted. For example, if a given dataset does not show a fixed effect estimate consistent with a tradeoff, and there is little random-effect variability relating the experimental condition effects to RT, then there would be few reasons to expect that a more elaborate model of the relation would reveal a tradeoff, unless more elaborate assumptions are adopted.

One difference between the models we have presented here, and that of Loeys et al., is that we specified both a fixed effect for RT on the probability of a response, as well as random-effect deviations from this fixed effect. The advantage of this fixed effect parameterization is that the population-level slope relating RT to accuracy can be examined directly. When there is an interaction between condition effects and the RT regression input, as in the example data analysis presented above, the relative magnitude of the interaction with respect to the general RT–accuracy relation can be assessed. This is in contrast to the random-effects only model specification, which furnishes the correlation of the random effect deviations (for either subjects or items), but it does not provide a slope parameter. In future work, it should be possible to extend the approach advocated by Loeys et al. to include both fixed and random effect RT–accuracy relations.

Table 6
Parameter estimates for the three models in the second simulation. The values in parentheses indicate standard errors.

Model	β_0	β_1	β_2
m0	−0.0816 (0.2454)	−1.6260 (0.3812)	−1.5295 (0.0909)
m1	−0.0679 (0.2655)	−2.1299 (0.5189)	−2.0749 (0.3593)
m2	−0.0593 (0.2762)	−2.2753 (0.5839)	−2.1520 (0.4006)

Table 7
Model comparison of the three models for the second simulation. The deviance scores (AIC and BIC) are smallest for model **m2**.

	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
m0	5	1408.53	1435.42	−699.27			
m1	6	1265.95	1298.22	−626.98	144.58	1	0.0000
m2	7	1255.23	1292.87	−620.62	12.72	1	0.0004

Based on the results presented here, we argue that it would be helpful in future studies using a binomial response to consider including response time as a fixed- and a random-effect input in a mixed effects regression model. Doing so can help uncover whether there are systematic relationships between reaction time and accuracy in a given dataset, both on the individual and group level, and whether this relationship plays a role in explaining differences between experimental conditions. Note also that to explore this relationship further, one may also model RT as a function of accuracy, or take a joint modeling approach as suggested by Loeys et al. (2011).

Appendix A. Code for simulation 1

First the data-generating function is parameterized to generate a random distribution of subjects. The size of the data set is set up, and the RTs and responses are sampled and assigned to a data frame. Also subject identifiers are assigned. Note that in this simulation, there is no experimental condition — it is simply attempting to simulate a negative relationship between accuracy and speed.

```
## Data generation

# Generating function
genData <- function(n){

    x1 <- rnorm(n,0,1)
    b0 <- rnorm(1,2,1)
    b1 <- rnorm(1,-b0,1)
    r1 <- rbinom(n,1,invlogit(b0+b1*x1))

    return(list(d=x1,r=r1))
}

# Parameters for the dataset
Nsbj <- 16
Nrsp <- 100

# Generate the response times and responses, and assign them to vectors
x   <- replicate(Nsbj, genData(Nrsp))

RT  <- unlist(x[seq(from=1,to=length(x),by=2)])
Rsp <- unlist(x[seq(from=2,to=length(x),by=2)])

# Set up the data frame
Sbj <- paste("s", rep(1:Nsbj, each=Nrsp), sep="")
d <- data.frame(Sbj, RT, Rsp)
```

In the modeling part of the simulation, three mixed model GLMs are set up.

```
## Modeling

# Fit the models and test
(m0 <- glmer(Rsp~RT+(1|Sbj), family=binomial(link="logit"), data=d))
(m1 <- glmer(Rsp~RT+(1|Sbj)+(0+RT|Sbj), family=binomial(link="logit"), data=d))
(m2 <- glmer(Rsp~RT+(1+RT|Sbj), family=binomial(link="logit"), data=d))
anova(m0, m1, m2)
```

Appendix B. Relationship between the *lmer* and *JAGS* models

Although the glmer model that we propose (which we will term here an "individual" model) and the model proposed by Loeys et al. (2011) are both mixed effects models, there are a number of important differences. The Loeys et al. model is fit within a Bayesian inference scheme using BUGS or JAGS. Most importantly, there is a joint model for both the RT and the response outcome, with a shared covariance matrix. The off-diagonal elements of the covariance matrix capture the correlation between the random effect intercepts of the RT with the random effect intercepts of the response. In the model they proposed, this correlation holds over participants (or items), and not within participants.

In contrast, the *glmer* model for the response outcome we propose is estimated not with simulation of distributions of the parameters, but rather point estimates of the statistics. Instead of a joint model with a shared covariance matrix, we include RT as a regression input for the response. Importantly, the trial-level RT regression input captures the correlation between the response and the RT within a given subject (or within a given item). It does not capture the correlation of intercepts over participants.

The last point is significant for evaluating whether one or the other model is appropriate for a given dataset, so it is perhaps worth outlining further. Fig. 5 shows a schematic illustration of two different types of correlations between accuracy and RT. Fig. 5A shows five participants with the same average accuracy, but a range of different average RTs, from low to high (note that the RTs are centered at zero). Within each participant, however, the slope of the curve relating accuracy to RT is positive. This illustrates how each participant can have a correlation between accuracy and RT, but at the same time across the participants there is no systematic relationship between *average* accuracy and *average* RT.

In contrast, Fig. 5B shows five participants with increasing average accuracy (the intercept of each line runs from low to high accuracy) as a function of average RT. Within each participant, however, there is no systematic relationship between accuracy and RT at the level of individual trials — the slope of each line is zero. Fig. 5B shows that it is possible to have a relationship between accuracy and RT at the level of averages, even if there is no relationship within each subject at the level of the trials. The relationships shown in Fig. 5A and B illustrate two different situations both well suited for mixed effects models because of the ability of these models to capture hierarchies. Note that this figure is an idealization, as hierarchical relationships like this can

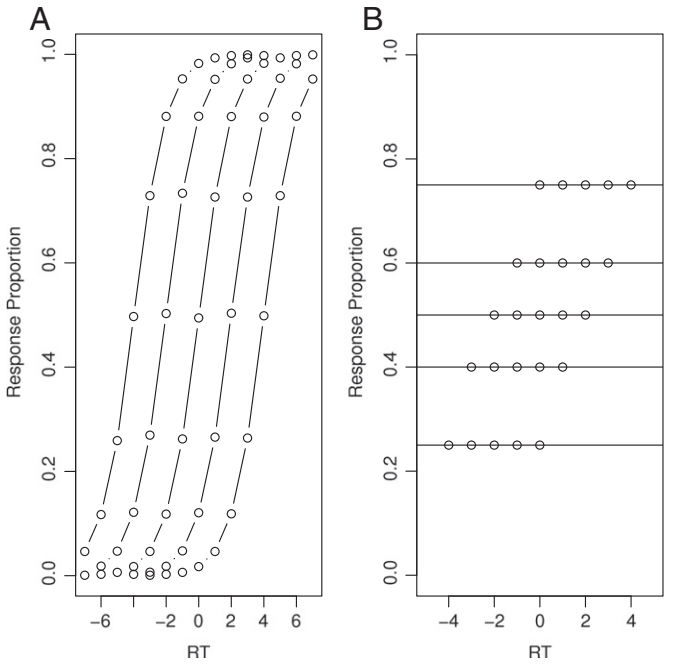

Fig. 5. Schematic illustration of (A) a correlation between accuracy and RT within participants (over trials), but no correlation over participants, and (B) a correlation between average accuracy and average RT over participants but no correlation within participants over trials.

hold for items as well as subjects, and there can be considerable heterogeneity among subjects (or items) in a dataset from an experiment.

We set up a simulation corresponding to the scenarios shown in Fig. 5 to illustrate the relationship between the mixed effects model we are proposing in this paper to the model proposed by Loeys et al. (2011). The section following shows the R code for generating the two samples, a JAGS model like the Loeys et al. (2011) approach for the joint distribution of the response and the RT, and finally example *glmer* mixed effects models for the two cases. In these two simulations, we concentrate on the comparison of trial-level versus subject-level accuracy–RT relations with respect to intercepts. Unlike simulation 1 described above, we do not simulate a correlation between the intercept and slope, mainly to simplify the simulations.

This simulation was arranged using a multivariate normal distribution, more in line with the work of Loeys et al. (2011), but similar in structure to the simulation 1 in the previous section. As in simulation 1, we are attempting to simulate a relationship between accuracy and RT for a sample of subjects, but without any contrast between experimental conditions or other variables. It is convenient to use the multivariate normal distribution here to illustrate two situations. In the first (genData1), the two means characterizing the bivariate normal are themselves not correlated with each other — they are two draws from $N(0,1)$. However, the covariance matrix is parameterized so that there is a high positive correlation between the two distributions. The correlation parameters *sigma*[1,2] and *sigma*[2,1] are set to 0.9.

In the second situation *genData2*, the two means are arranged to be dependent on each other. The mean of the second variable of the bivariate normal distribution is drawn from a normal distribution with the mean set to be the same as the draw from the first normal distribution. The variance of the second variable is set to be half that of the first (i.e., *rnorm*(1,*mn*[1],0.5) in order to simulate a relatively high (positive) correlation. To arrange a negative correlation, the second mean could be defined as *rnorm*(1,*mn*[1],0.5), for example.

Given these two data-generating functions, two simulated datasets are arranged. The first has a high correlation between accuracy and RT within participants at the level of trials, but a low correlation of the average accuracy and average RT across participants. The second has the opposite structure: a low correlation within participants at trial level, but a high correlation across participants.

```
# ##########################################################################
# Data generating function
# Version 1: Correlation within trials
# ##########################################################################


genData1 <- function(n){

    mn            <- rnorm(2,0,1)                    # Independent means


    sigma          <- diag(length(mn))
    sigma[1,2]     <- 0.9                            # High correlation
    sigma[2,1]     <- sigma[1,2]


    rsp            <- rmvnorm(n, mean = mn, sigma = sigma, method=c("eigen"))


    x1             <- rsp[,1]
    r1             <- rbinom(n, 1, invlogit(0 + rsp[,2]))

    return(list(d=x1,r=r1))
}
```

```
# ##########################################################################
# Data generating function
# Version 2: Correlation of means over sbj
# ##########################################################################

genData2 <- function(n){


    mn            <- rnorm(1,0,1)
    mn[2]         <- rnorm(1,mn[1],0.5)            # Dependent means


    sigma          <- diag(length(mn))
    sigma[1,2]     <- 0.1                           # Low correlation
    sigma[2,1]     <- sigma[1,2]


    rsp            <- rmvnorm(n, mean = mn, sigma = sigma, method=c("eigen"))


    x1             <- rsp[,1]
    r1             <- rbinom(n, 1, invlogit(0 + rsp[,2]))

    return(list(d=x1,r=r1))
}




    # Parameters for the datasets
    Nsbj <- 16
    Nrsp <- 100




    # Generate the response times and responses, and assign them to vectors
    x   <- replicate(Nsbj, genData1(Nrsp))
    RT  <- unlist(x[seq(from=1,to=lengt (x),by=2)])
    Rsp <- unlist(x[seq(from=2,to=length(x),by=2)])


    # Set up the data frame
    Sbj <- paste("s", rep(1:Nsbj, each=Nrsp), sep="")
    d1 <- data.frame(Sbj, RT, Rsp)




    # Generate the second dataset
    x   <- replicate(Nsbj, genData2(Nrsp))
    RT  <- unlist(x[seq(from=1,to=length(x),by=2)])
    Rsp <- unlist(x[seq(from=2,to=length(x),by=2)])


    # Set up the data frame for the second dataset
    Sbj <- paste("s", rep(1:Nsbj, each=Nrsp), sep="")
    d2 <- data.frame(Sbj, RT, Rsp)
```

The Loeys–JAGS model is set up as a joint model for the simulated RT and response outcomes. The syntax of the JAGS model is similar to

that of R, but in fact, the JAGS code is run via a program external to R, using the package *rjags*.

```
model{

for (k in 1:length(rt)){

    rt[k] ~ dnorm( y1.hat[k], tau.y1 )
    y1.hat[k] <- alpha0 + a1[sbj[k],1]

    rsp[k] ~ dbern( p[k] )
    logit(p[k]) <- beta0 + a1[sbj[k],2]

}

# Priors at the trial level
tau.y1 ~ dgamma(0.001,0.001)
sigma.y1 <- sqrt(1/tau.y1)

# Priors for random effects
for (k in 1:K){

    a1[k,1] <- B[k,1]
    a1[k,2] <- B[k,2]

    B[k,1:2] ~ dmnorm(B.hat[k,1:2], tau.B[1:2,1:2])

    B.hat[k,1]        <- alpha0
    B.hat[k,2]        <- beta0

}

# Priors for rt-model parameters
alpha0 ~ dnorm(0, 0.001)

# Priors for rsp-model parameters
beta0 ~ dnorm(0, 0.00001)

# Priors for the covariance matrix
tau.B[1:2,1:2] <- inverse(sigma.B[,])
sigma.B[1,1]   <- pow(sigma.a1, 2)
sigma.a1 ~ dunif (0, 100)
sigma.B[2,2]   <- pow(sigma.a2, 2)
sigma.a2 ~ dunif (0, 100)
sigma.B[1,2]   <- rho*sigma.a1*sigma.a2
sigma.B[2,1]   <- sigma.B[1,2]
rho ~ dunif (-1, 1)

}
```

Within R, after loading the package “rjags”, the simulation is initialized and run via *jags.model* and *coda.samples*. The model that was defined above has been saved to a text file named “model.txt”. The summary of the output of tcoda.samples will show quantiles of the posterior distribution of the parameters listed in “variable.names”.

Below, the samples are drawn for the model using the data from the first simulated dataset (*d1*). The samples can be obtained similarly for the second simulated dataset (*d2*).

```
library(rjags)

K = length(unique(d$Sbj))
N = nrow(d1)

rsp = d1$Rsp
sbj = as.numeric(d1$Sbj)
rt  = d1$RT

pathtomodel <- "model.txt"

bm.sim <- jags.model( file=pathtomodel,...
        data=list('K'=K,'rsp'=rsp,'rt'=rt,'sbj'=sbj),...
                    n.chains=4,...
                    n.adapt=5000,...
                    quiet=FALSE)

bm.smp <- coda.samples( bm.sim,...
variable.names=c("a1","alpha0","beta0","sigma.y1", "rho"),...
n.iter=20000,...

thin=10)

summary(bm.smp)
```

The glmer models for the response are set up as in the previous Appendix Section A. To simplify matters, we restrict our attention to models with (*m1*) and without (*m0*) *RT* as a random effect regression input.

```
# Fit the models and test
(m0 <- glmer(Rsp~1+(1|Sbj), family=binomial(link="logit"), data=d1))
(m1 <- glmer(Rsp~1+(1|Sbj)+(0+RT|Sbj), family=binomial(link="logit"), data=d1))
anova(m0, m1)
```

For the comparison to the Loeys–JAGS model, a model with average RT as the regression input can be formulated as follows:

```
# Add a column containing the mean RT of each participant
library(doBy)

d$MnRT <- summaryBy(RT~Sbj,...
    FUN=mean,...
                    data=d1,...
                    full.dimension=TRUE,...
                    order=FALSE)$RT.mean

# Fit the model with mean RT rather than RT
(m3 <- glmer(Rsp~MnRT+(1|Sbj), family="binomial", data=d1))
```

Appendix B.1. Results

For the dataset representing the situation shown in Fig. 5A, the *glmer* model with the random effect RT regression input (**m1**) was clearly better than the model without this regression input (**m0**):

	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
m0	3	2187.02	2203.15	−1090.51			
m1	4	2010.89	2032.40	−1001.44	178.13	1	0.0000

For the dataset corresponding to Fig. 5B, there was no strong evidence for one *glmer* model over the other:

	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
m0	3	2206.10	2222.23	−1100.05			
m1	4	2205.27	2226.78	−1098.63	2.83	1	0.0923

Note we cannot formulate exactly the model that Loeys et al. use with *glmer* because it is not set up to estimate a shared covariance matrix between jointly estimated equations. This is one advantage of the JAGS (or BUGS) language that flexible models like this can be formulated. However, we can approximate the relationship they model by including *mean RT* as a regression input for the binary response. The sign of the coefficient for this regression input will provide some indication of the type of correlation to be expected in the Loeys–JAGS model. We used this approach for the two simulation datasets. Please see Loeys et al. (2011) for a more complete comparison.

For the example dataset 1, with the correlation *within* but not *over* subjects, the *glmer*-estimated coefficient for the average RT was 0.020, SE = 0.191, z = 0.104, and p = 0.917. The magnitude of this coefficient is small relative to its standard error, indicating that there is little systematic relationship between the average RT and the response. We already know from the results presented above that in this dataset the single trial RT was an effective regression input of the single-trial response. This suggests that this dataset would be better modeled with the *glmer* approach we advocate here, in order to explore the relationship between accuracy and RT within individual subjects, rather than across subjects.

The joint Loeys–JAGS model for this dataset was estimated using four chains of 20 K samples following a burn-in of 5 K samples. Fig. 6A shows a plot of the mean random effects across the four chains. As expected from the simulation parameters, the estimated correlation of the random effects of subject for the RT and the response was near zero for this dataset: $\rho = 0.0214$, and the confidence interval included zero: 95% CI = −0.483, 0.521. The chains for all parameters appeared to have mixed well based on plots of the samples and other diagnostic plots, and all the parameter densities were unimodal with relatively symmetric distributions.

For the example dataset 2, with the simulated positive relationship *over* averages (but not *within* participants), the *glmer*-estimated coefficient for average RT corresponded to 0.775, SE = 0.080, z = 9.659, and $p < 2e-16$. The positive value of the coefficient, along with its relatively large magnitude compared to its standard error, provides an important clue that in this dataset there is a positive relationship between average RT and the response. This is the type of relation that the JAGS joint approach is set up to model.

The same JAGS model as the first dataset was estimated with similar simulation parameters. Fig. 6B shows the mean random effects across the four chains. As the plot shows, there was a strong positive correlation of the random effect subject intercepts for the RT and the response for this dataset, as expected from the simulation: $\rho = 0.957$, and the confidence interval excluded zero 95% CI = 0.801, 0.996. As with the previous JAGS model, the chains appeared to have mixed well, and most all parameter estimates had symmetric, unimodal distributions. Notably, however, the parameter estimate for ρ was somewhat skewed because its estimate is near a boundary, but nevertheless all chains converged to a similar value.

The two simulations in this appendix, along with the associated different statistical models, have shown the relative effectiveness of the joint modeling approach and the individual modeling approaches. In a situation where the accuracy–RT relation holds mainly over averages across participants (or items), the joint approach may be a better

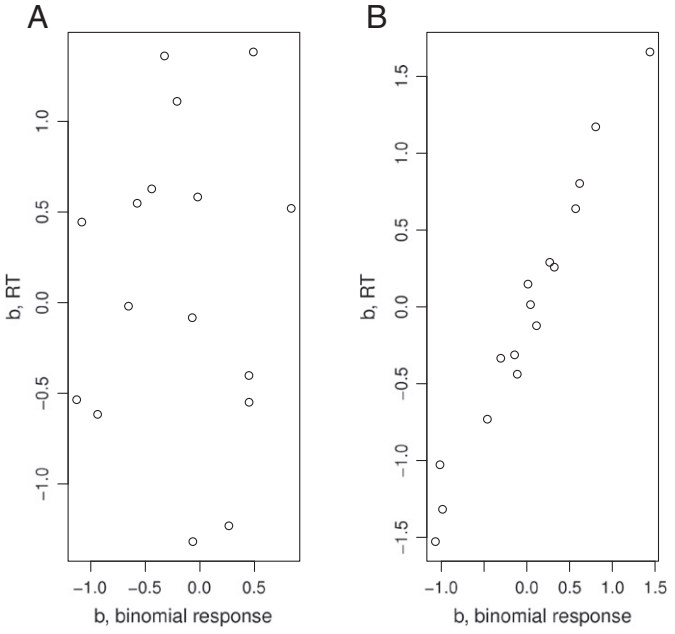

Fig. 6. Plots of (A) the relationship between accuracy and RT for the first simulated dataset, (with no correlation over participants), and (B) the second simulated dataset, with a strong positive correlation between average accuracy and average RT over participants but no correlation within participants over trials. The individual points in both plots are the estimated random effects taken from the same JAGS model.

approach, due to the flexibility of the JAGS modeling language. In a situation where the relation holds mainly within and not across participants (or items), the individual approach may be better suited for exploring variability within a dataset. Note that it is currently not possible to combine the joint and individual approaches within JAGS by including RT itself as a regression input for the response (and vice versa, use the response as a regression input for RT), as JAGS prohibits simultaneous equations parameterized this way. Please note that there are other potential approaches to this modeling problem. Future work could explore other analysis approaches, including (as suggested to us by an anonymous reviewer), approaches that allow a multivariate response (MCMCglmm; see Hadfield, 2010).

Appendix C. Code for simulation 2

```
## Data generation
# Function to generate n points for one subject
genData <- function(n){

    x1 <- rnorm(n,0,1)      # Range of (standardized) RTs
    x2 <- rbinom(n,1,0.5)   # Condition design matrix
    b0 <- rnorm(1,0,1)      # Random effect intercept deviation
    b1 <- rnorm(1,-2,1)     # Random effect RT slope deviation
    b2 <- rnorm(1,b1,1)     # Random effect condition deviation
    r1 <- rbinom(n,1,invlogit( b0 + b1*x1 + b2*x2)) # Generate data

    return(list(d=x1,d2=x2,r=r1))

}

# Parameters for the dataset
Nsbj <- 16
Nrsp <- 100
Ncnd <- 2

# Generate the response times, conditions, and responses; and assign to vectors
x   <- replicate(Nsbj, genData(Nrsp))
RT  <- unlist(x[seq(from=1,to=length(x),by=3)])
Cnd <- unlist(x[seq(from=2,to=length(x),by=3)])
Rsp <- unlist(x[seq(from=3,to=length(x),by=3)])

# Set up the data frame
Sbj <- paste("s", rep(1:Nsbj, each=Nrsp), sep="")
d <- data.frame(Sbj, RT, Cnd, Rsp)

## Modeling
# Fit the models and test
(m0 <- glmer(Rsp ~ Cnd+RT + (1|Sbj),
             family=binomial(link="logit"),
             data=d))

(m1 <- glmer(Rsp ~ Cnd+RT +
             (1|Sbj) + (0+Cnd|Sbj) + (0+RT|Sbj),
             family=binomial(link="logit"),
             data=d))

(m2 <- glmer(Rsp ~ Cnd*RT + (1|Sbj) + (0+RT+Cnd|Sbj),
             family=binomial(link="logit"),
             data=d))

anova(m0,m1,m2)
```

References

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*(4), 390–412.

Bates, D. M. (2010). *lme4: Mixed-effects modeling with R.* New York: Springer (Prepublication version at: http://lme4.r-forge.r-project.org/book/).

Boersma, P., & Weenink, D. (2005). *Praat: Doing phonetics by computer [computer program].* (Version 5.2.17).

Davis, C. J., & Perea, M. (2005). BuscaPalabras: A program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indices in Spanish. *Behavior Research Methods*, *37*(4), 665–671.

Fitts, P. M. (1966). Cognitive aspects of information processing: III. Set for speed versus accuracy. *Journal of Experimental Psychology*, *71*, 849–857.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* New York: Cambridge University Press.

Garret, H. E. (1922). *A study of the relation of accuracy to speed.* Archives of psychology, New York, No. 56.

Hick, W. E. (1952). On the rate of gain of information. *Quarterly Journal of Experimental Psychology*, *4*, 11–26.

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software*, *33*, 1–22 (http://www.jstatsoft.org/v33/i02/).

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446.

Jaeger, T. F., Graff, P., Croft, B., & Pontillo, D. (2011). Mixed effect models for genetic and areal dependencies in linguistic typology: Commentary on Atkinson. *Linguistic Typology*, *15*(2), 281–319.

Kliegl, R., Masson, M. E. J., & Richter, E. M. (2010a). A linear mixed model analysis of masked repetition priming. *Visual Cognition*, *18*, 655–681.

Kliegl, R., Wei, P., Dambacher, M., Yan, M., & Zhou, X. (2010b). Experimental effects and individual differences in linear mixed models: Estimating the relationship between spatial, object, and attraction effects in visual attention. *Frontiers in Psychology*, *1*, 238.

Kroll, J., & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language*, *33*, 149–174.

Loeys, T., Rosseel, Y., & Baten, K. (2011). A joint modeling approach for reaction time and accuracy in psycholinguistic experiments. *Psychometrika*, *76*, 487–503.

Ollman, R. (1966). Fast guess in choice reaction time. *Psychonomic Science*, *6*, 155–156.

Pachella, R. G., & Fisher, D. F. (1969). Effects of stimulus degradation and similarity on the tradeoff between speed and accuracy in absolute judgments. *Journal of Experimental Psychology*(81), 7–9.

Pachella, R. G., & Pew, R. W. (1968). Speed–accuracy tradeoff in reaction time: Effect of discrete criterion times. *Journal of Experimental Psychology*, *76*, 19–24.

Perea, M., Urkia, M., Davis, C. J., Agirre, A., Laseka, E., & Carreiras, M. (2006). E-Hitz: A word frequency list and a program for deriving psycholinguistic statistics in an agglutinative language (Basque). *Behavior Research Methods*, *38*(4), 610–615.

Pew, R. W. (1969). The speed–accuracy operating characteristic. *Acta Psychologica*, *30*, 16–26.

Pinheiro, J. C., & Bates, D. M. (2000). Mixed effects models in S and S-PLUS. *Statistics and Computing Series*. New York, NY: Springer-Verlag.

Rabbitt, P. (1967). Time to detect errors as a function of factors affecting choice–response time. *Acta Psychologica*, *27*, 131–142.

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

Quené, & van den Bergh (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, *59*, 413–425.

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

Ratcliff, R. (1985). Theoretical interpretations of speed and accuracy of positive and negative responses. *Psychological Review*, *92*, 212–225.

Ratcliff, R., & Hacker, M. J. (1981). Speed and accuracy of same and different responses in perceptual matching. *Perception & Psychophysics*, *30*, 303–307.

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

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A diffusion model account of the lexical decision task. *Psychological Review*, *111*, 159–182.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*, 873–922.

Reed, A. (1973). Speed-accuracy tradeoff in recognition memory. *Science*, *18*, 574–576.

Reed, A. (1976). List length and the time-course of recognition in immediate memory. *Memory and Cognition*, *4*, 16–30.

Schouten, J. F., & Bekker, J. A. M. (1967). Reaction time and accuracy. *Acta Psychologica*, *27*. (pp. 143–153).

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.

Tyers, F. M., Sánchez-Martínez, F., Ortiz-Rojas, S., & Forcada, M. L. (2010). Free/open-source resources in the Apertium platform for machine translation research and development. *The Prague Bulletin of Mathematical Linguistics*, *93*, 67–76.

Van der Linden, W. (2007). A hierarchical framework for speed and accuracy on test items. *Psychometrika*, *72*(3), 287–308.

Wickelgren, W. A. (1977). Speed–accuracy tradeoff and information processing dynamics. *Acta Psychologica*, *41*, 67–85.

Woodworth, R. S. (1899). Accuracy of voluntary movement. *Psychological Review Monographs*, *27–54*.