This is a preprint of an article whose final and definitive form was published in *Visual Cognition*,

Kliegl, Reinhold , Masson, Michael E. J. andRichter, Eike M.(2010) 'A linear mixed model analysis of masked repetition priming', Visual Cognition, 18: 5, 655 — 681, First published on: 03 August 2009 (iFirst)

To link to this Article: DOI: 10.1080/13506280902986058 URL: http://dx.doi.org/10.1080/13506280902986058

A linear mixed model analysis of masked repetition priming

Reinhold Kliegl

Department of Psychology, University of Potsdam, Germany

Michael E. J. Masson

Department of Psychology, University of Victoria, Canada

Eike M. Richter

Department of Psychology, University of Potsdam, Germany

We examined individual differences in masked repetition priming by re-analyzing item-level response-time (RT) data from three experiments. Using a linear mixed model (LMM) with subjects and items specified as crossed random factors, the originally reported priming and word-frequency effects were recovered. In the same LMM, we estimated parameters describing the distributions of these effects across subjects. Subjects' frequency and priming effects correlated positively with each other and negatively with mean RT. These correlation estimates, however, emerged only with a reciprocal transformation of RT (i.e., -1/RT), justified on the basis of distributional analyses. Different correlations, some with opposite sign, were obtained (1) for untransformed or logarithmic RTs or (2) when correlations were computed using within-subject analyses. We discuss the relevance of the new results for accounts of masked priming, implications of applying RT transformations, and the use of LMMs as a tool for the joint analysis of experimental effects and associated individual differences.

How does an individual's mean response speed relate to that person's effect size in response to an experimental manipulation in a cognitive task? Somewhat surprisingly, there is no clear answer to this question. Even for well-studied experimental effects such as the relation between word frequency and masked repetition priming (Forster & Davis, 1984; Forster, Mohan, & Hector, 2003), we do not know whether fast responders show larger or smaller frequency or priming effects and whether frequency and priming effects correlate positively or negatively with each other. Here, we demonstrate how such individual differences in experimental effects and their correlations can be estimated simultaneously with the genuine effects of the experimental manipulation. Specifically, we show how individual differences, presumably present in all psychological experiments, can be included in the analysis of experimental effects by replacing traditional repeated-measures analyses of variance (rmANOVA) with a linear mixed model analysis (LMM). In a re-analysis of published data, we show (1) that correlations based on difference scores computed separately for each subject (i.e., within-subject analyses) are inferior to estimating such correlations in a LMM and (2) that the strength and even the sign of such correlations depend strongly on the metric one chooses for reaction times (RTs).

CORRELATIONS OF EXPERIMENTAL EFFECTS

Is the variability between subjects, typically seen when assessing RT effects, meaningfully linked to fundamentally important features of the cognitive architecture? For example, there may be a systematic relation between a subject's mean response speed and the size of the effect of a manipulation on that subject's RT. A relationship of this type may support conclusions about the relative speed with which separable cognitive operations are completed.

Correspondence: Reinhold Kliegl, Department of Psychology, University of Potsdam, Karl-Liebknecht-Str. 24-25, 14465 Potsdam-Golm, Germany. Email: kliegl@uni-potsdam.de

The research was initiated during Michael Masson's residence as a guest professor at the Interdisciplinary Center for

Cognitive Studies at the University of Potsdam. He was supported in part by a discovery grant from the Natural Sciences and Engineering Research Council of Canada. We are indebted to Douglas Bates for providing the *lme4* package in the R-project and for stimulating conversations about the interpretation of conditional means, formerly known as BLUPs, as well as their correlations. We are also very grateful to Sachiko Kinoshita for making available the data from the second experiment reported in Kinoshita (2006). Harald Baayen, Sachiko Kinoshita, Nicholas Lewin-Koh, John Maindonald, Wayne Murray, Klaus Oberauer, and reviewers commented on an earlier version of the manuscript. This research was supported by Deutsche Forschungsgemeinschaft (KL 955/6 and KL 955/8). Data and R-scripts are provided on request.

In support of this possibility, numerous studies examining RT distributions have shown that effects of independent variables can be particularly pronounced for slow responses (e.g., Balota, Yap, Cortese, & Watson, 2008; Ridderinkoff, 2002; Steinhauser & Hübner, 2008). Where this trend occurs, one might expect that slower subjects should generate larger effects of a manipulation. In this article, we reportsignificant correlations between average response speed, masked priming effects, and word frequency effects in a lexical decision task using a joint re-analysis of three experiments reported in Bodner and Masson (1997; Exp. 1 and 2) and Kinoshita (2006, Exp. 2). This analysis, however, can be applied to many psychological experiments. Therefore, before we turn to the specifics of our data set, we describe our approach from a general perspective.

We typically manipulate some independent variables within subjects to provide a powerful statistical test of effects. Subjects vary in the size of such effects and this variability is treated as error or noise in standard analysis of variance models. But usually this variability is also indicative of reliable individual differences in the experimental effects. A reasonable starting point for examining this possibility is to test whether there is a positive or a negative relationship between, for example, subjects' mean RT and their various raw experimental effect sizes. Simple introspection affords predictions for both positive and negative correlations.

For example, subjects who "take their time" in general might give an experimental effect a better chance to express itself. This is a well-known result from individual-differences research. For example, older adults are typically slower on many RT tasks than young adults and also show larger absolute effect sizes, yielding the typical age x task complexity interaction (e.g., as noted already by Birren, 1956). If we ignore age group, such a pattern of results translates into a positive correlation between mean RT and effect size across all subjects (i.e., the so-called ecological fallacy). Obviously, such a correlation is likely to exist also within homogeneous age groups on the basis of normal interindividual differences. Similarly, degrading a stimulus leads to longer RT and also to enhanced effects of factors such as semantic context (e.g., Becker & Killion, 1977; Borowsky & Besner, 1993).

From a different perspective, we might instead expect a negative correlation between mean RTs and masked-priming effect sizes. Subjects who are more skilled (faster) at identifying words might be able to more successfully encode a briefly presented word prime and could use that information to more efficiently process a subsequently presented target word. Alternatively, suppose that subjects differ in their degree of task engagement. Those who comply with the instruction to respond as fast as possible will have shorter RTs than subjects with a casual attitude towards the experiment. The latter subjects may be less likely to attend closely to the visual display and could therefore fail to encode information from the masked primes, leading to relatively weak priming effects. Under either of these two scenarios, subjects with shorter RTs could be more sensitive to differences in word frequency because they might base their responses on relatively little accumulated information (e.g., Wagenmakers, Ratcliff, Gomez, & McKoon, 2008) and the influence of word frequency may be especially strong at early stages of word processing. Importantly, in either event, mean RT should correlate negatively with all experimental effects, but the experimental effects should correlate positively among each other. Moreover, a positive correlation between two experimental effects in the absence of any correlation between either of those effects and mean RT would strongly suggest an architecturally relevant relationship between the two effects. The LMM analysis we present here allows us to test these competing possibilities.

APPLICATION TO EFFECTS OF MASKED REPETITION PRIMING AND WORD FREQUENCY

In this article, we re-analyze effects of masked repetition priming and word frequency with LMMs. In experimental research, statistical analyses emphasize the significance of main effects and their interactions—so called fixed effects. In the reports by Bodner and Masson (2001) and Kinoshita (2006), the hypothesis was that masked repetition effects might be larger for low- than for high-frequency words. Initially, the rationale for this proposal was that in studies of long-term repetition priming, low-frequency words produce reliably more repetition priming than high-frequency words (e.g., Forster & Davis, 1984; Jacoby & Dallas, 1981). Bodner and Masson (2001) proposed that masked repetition priming and long-term priming may have a common basis in a form of memory for the processing of the prime event and should therefore operate according to a common set of principles. Thus, the well established interaction between word frequency and repetition seen in long-term priming was expected to appear with masked priming as well. Although Bodner and Masson (2001) obtained such an interaction, Bodner and Masson (1997), using a weaker manipulation of word frequency, failed to do so in two separate experiments. The

difference between those two experiments was the visual format in which targets items were presented (i.e., in normal uppercase or alternating case). In a related study, Kinoshita (2006) was able to produce an interaction between frequency and masked priming by ensuring that even the low-frequency word targets were familiar to subjects. Kinoshita's reasoning was that the low-frequency words used in earlier studies (including her Experiment 1) were of low semantic familiarity and therefore were not stably represented in the mental lexicon. Consequently, these low-frequency items were not capable of reliably activating a lexical representation when presented as a masked prime. By using familiar low-frequency words, it was expected that these items would successfully prime the lexicon when presented as masked primes, leading to full repetition priming effects and generating an interaction between frequency and priming. The LMM that we applied to the original data from these three experiments (Bodner & Masson, 1997, Exp. 1 and 2; Kinoshita, 2006, Exp. 2) was expected to lead to the same conclusions as the original reports as far as the significance of main effects and interactions is concerned.

Our emphasis in this article is on correlations between (1) mean RT, (2) size of priming effect (i.e., the difference between RTs in unrelated and repetition prime conditions), and (3) size of frequency effect (i.e., the difference between RTs to low- and high-frequency word targets) across subjects. In individual differences research, these correlations are typically computed in separate analyses of mean RTs and difference scores based on individual subjects' data. Recent work has shown that the reliability of certain effects, particularly of semantic priming, is surprisingly low (Stolz, Besner, & Carr, 2005), so it is critical to take into account the reliability of measures when examining individual differences. In contrast to such a within-subject analysis, a LMM estimates parameters representing the variances (standard deviations) and covariances (correlations) of these effects across subjects (i.e., the variance component parameters) simultaneously with the fixed effects. The LMM parameters afford a better prediction of subjects' individual mean RTs as well as of their frequency and priming effects and correlations than is accomplished with a within-subject analysis because they take into account between-subject differences in reliability of mean RTs as well as of frequency and priming effects (i.e., the predictions are a type of shrinkage estimate; Faraway, 2006).

As it turns out, there is another very critical issue requiring attention in the analyses of individual differences in experimental effects. Correlations between effect sizes depend strongly on distributional properties of the dependent variable. RT distributions, for example, typically exhibit a positive skew, violating the normal distribution assumption. Such violations can be corrected with a suitable power transformation, using, for example, the Box-Cox procedure to estimate the optimal power coefficient (Box & Cox, 1964). Typically, in the case of RTs, scientists apply a log transformation or take the reciprocal of standard RTs. The former transformation moves statistical inferences into a multiplicative frame, whereas reciprocal RTs afford an interpretation of effects in terms of rate or speed rather than time. Obviously, these transformations preserve the ordinal relation of means, so they do not change the direction of effects. Actually, they rarely even affect the significance of main effects. Matters are not straightforward, however, for their influence on interactions (e.g., Loftus, 2002). For instance, a log transformation will render a significant interaction for standard RTs insignificant when similar proportional differences exist between pairs of means but will induce a significant subadditive interaction in log RTs when a pure main-effect pattern holds for simple RT. Moreover, as we demonstrate here with separate LMMs for untransformed, log-transformed, and reciprocal RTs, the choice of transformation may even change the sign of the correlation between effects.

In summary, we combined and reanalyzed the data from three published experiments on masked repetition priming. In each experiment, the key independent variables were relatedness of prime-target pairs and target frequency. We replicated the ANOVA-based inferences of the original publications for untransformed, log-transformed, and reciprocal RTs, and also estimated the variances and correlations associated with these effects across subjects. We will show that these estimated correlations yield a much clearer picture than correlations computed directly from the observed RTs of individual subjects (i.e., within-subject estimates). Counter to established practice, correct statistical inference about such correlations depends critically on a transformation of RTs that establishes compliance with distributional assumptions.

Method

Subjects. Results are reported for 72 students, 24 having participated each in Experiments 1 and 2a of Bodner and Masson (1997) and Experiment 2 of Kinoshita (2006).

Materials and procedure. In the Bodner and Masson (1997) experiments, subjects were presented a sequence of 204 masked priming trials in a lexical decision task. Of these, 96 were critical trials that presented a word target. Half of the word targets were low frequency and half were high frequency. Half of the word targets of each frequency were preceded by an identity prime appearing in lowercase letters (duration: 60 ms) and the other half

were preceded by an unrelated word prime. Assignment of items to prime conditions was counterbalanced across subjects. Targets appeared in uppercase letters in Experiment 1 and in alternating case in Experiment 2a. Data from Kinoshita's (2006) second experiment were also available in the unaggregated format required for the LMM analyses. Subjects were presented a sequence of 216 masked priming trials in a lexical decision task. Of these, 96 were critical trials that presented a word target. Half of the word targets were low frequency and half were high frequency. Half of the word targets of each frequency were preceded by an identity prime appearing in lowercase letters and the other half were preceded by an unrelated word prime. Assignment of items to prime conditions was counterbalanced across subjects. The critical feature of Experiment 3 was the selection of low-frequency words that were of high familiarity, that is, a minimum familiarity rating of 490 on a scale of 100–700 based on the MRC Psycholinguistic Database (Coltheart, 1981). Each trial began with a forward mask (a row of Xs) for 500 ms. Prime duration was 60 ms in Bodner and Masson (1997) and 53 ms in Kinoshita (2006). Subjects classified each target as a word or a nonword. Reaction time and response accuracy were measured on each trial.

Data screening. The following analyses are based on RTs from correct trials with high- and low-frequency target words following identity and unrelated masked primes. Excluding incorrect trials and the two shortest response latencies (i.e., < 250 ms) left us with 4182 of 4608 RTs (i.e., 91%) from Bodner and Masson (1997) and with 2199 of 2304 RTs (95%) from Kinoshita (2006). There were statistically reliable effects associated with errors in a generalized linear mixed model (GLMM); the effects went in the same direction as RTs, that is opposite to a potential speed-accuracy tradeoff.

Analysis software. We used the *lmer* program of the *lme4* package (Bates, Maechler, & Dai, 2009) for estimating fixed effects and variance/covariance component parameters of the LMM (see Bates, 2008a, 2008b, for technical background). This package and many others (e.g., we extensively used *lattice,* Sarkar, 2008, *reshape,* Wickham, 2007, and *ggplot2,* Wickham, 2009) are supplied in the *R* system for statistical computing (version 2.8.1 R Development Core Team, 2009) under the GNU General Public License (Version 2, June 1991).

Fixed effects. We coded priming and frequency effects as +.5/-.5 contrasts (i.e., unrelated - repetition primes, low - high frequency words) and the two contrasts associated with the three experiments as two orthogonal Helmert contrasts (C1: Bodner-Masson-Exp 1 vs. Bodner-Masson-Exp 2a; C2: both BM-Exps vs. Kinoshita-Exp).[1]

Random factors and variance/covariance component parameters. Subjects and words are specified as random factors, varying in mean RTs. We also assume that subjects vary reliably in frequency and priming effects. The LMM assumes that words' mean RTs as well as subjects' mean RTs, priming effects, and frequency effects are normally distributed around the respective fixed effects (i.e., the grand mean RT, the mean difference between unrelated and repetition primes, and the mean difference between low- and high-frequency words). This specification yields six variance/covariance component parameters for subjects and one variance component parameter for words (see Baayen, 2008, and Baayen, Davidson, & Bates, 2008, for discussion of replacing F1/F2-ANOVA with LMM). Finally, the LMM also estimates the residual variance.

Results

Figure 1 displays the priming x frequency interaction for each of the three experiments (columns) for untransformed RT, log-transformed RT, and -1/RT (rows). We multiplied reciprocal scores by minus 1 to maintain the direction of effects compatible for the three variants, effectively converting speed into "rate of slowing". The pattern of means reveals larger priming effects for low-frequency than for high-frequency words in Experiment 3, but no support for this interaction in Experiments 1 and 2. Standard and transformed RTs afford the same interpretation. Tables 1 and 2 display parameter estimates for fixed effects and variance/covariance components, respectively.

Fixed effects. The fixed-effect estimates of untransformed RT, log RTs, and reciprocal RTs are listed in separate columns of Table 1. Our criterion for significance was a coefficient magnitude of at least two standard errors (i.e., absolute t values > 2). The degrees of freedom for t-values are not known exactly for a LMM. Given the large number of observations in our analyses, however, the t distribution has converged, for all practical purposes, to the standard normal distribution. In this case the 2-SE criterion is close to the conventional two-tailed 5% level of significance (e.g., Baayen et al., 2008, Note 1).[2] In agreement with the visual impression conveyed by Figure 1, raw

[1]It would certainly be in the spirit of LMM to use continuous frequency values rather than two extreme frequency categories. However, we prefer to respect the design choices of the original publications for ease of comparison. For continuous, usually log-

transformed, frequencies, the fixed effect represents the linear regression slope for RT on word frequency. The random effect of frequency represents the between-subject variance in linear regression slopes. Linear, quadratic and even cubic fixed effects of log frequency have been reported for single-fixation durations in reading (e.g., Kliegl, 2007).

[2]There is also the option to use Markov Chain Monte Carlo (MCMC) methods to generate a sample from the posterior distribution of the parameters of a fitted model and determine the approximate highest 95% posterior density (HPD) interval for the coefficients in this sample. In our experience, typically involving large data sets like the present one, inferences based on HPD intervals have been overwhelmingly consistent with the $t > 2$ criterion.

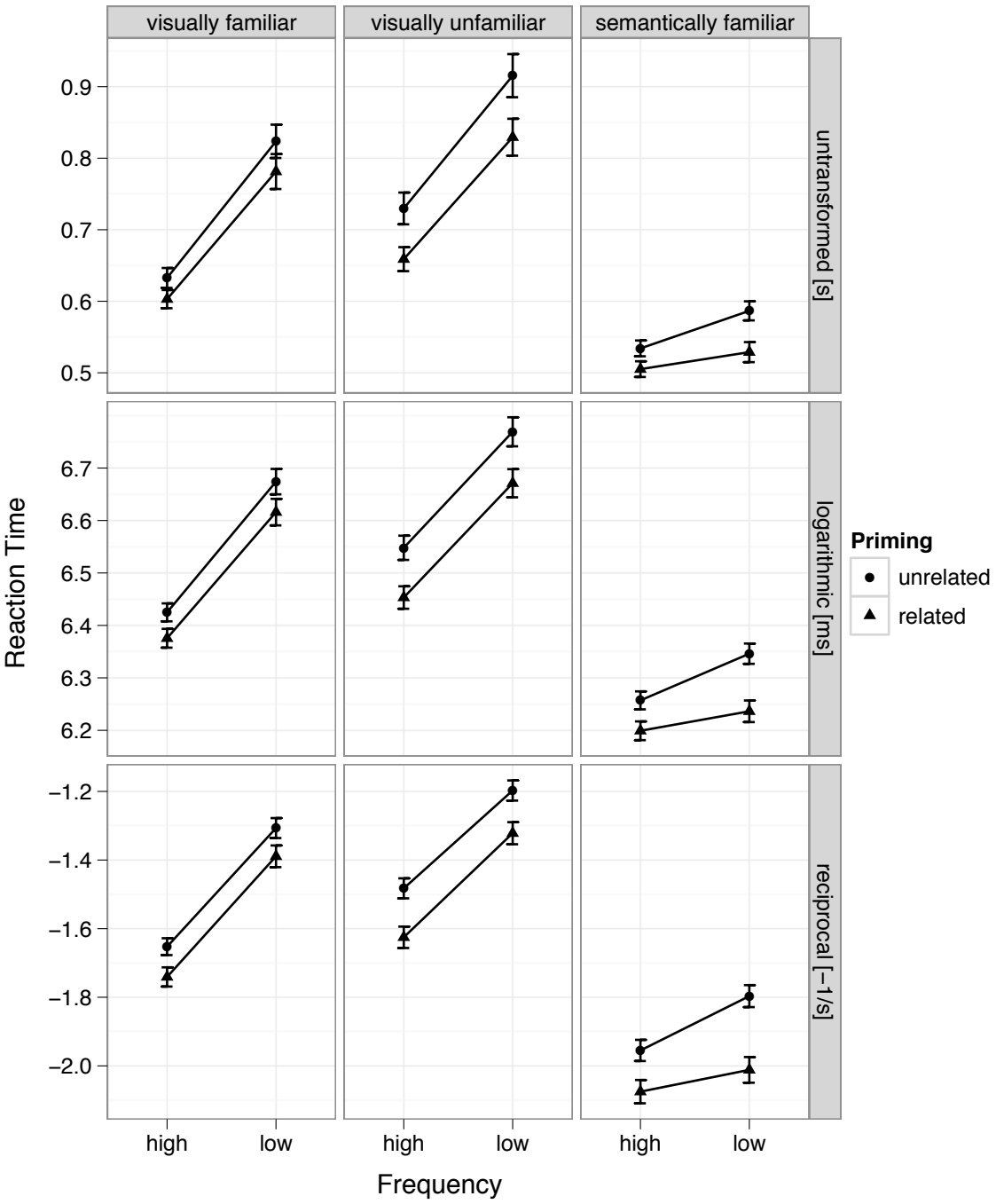

Figure 1. Each row shows the frequency by priming interaction for (a) Bodner and Masson (1997, Exp. 1), (b) Bodner and Masson (1997, Exp. 2a), and (c) Kinoshita (2006, Exp. 2). Effects are displayed for untransformed RT (top row), logarithmic RT (middle row), and reciprocal RT (i.e., -1/RT, bottom row). Error bars represent 95% confidence intervals for cell means (i.e., they are not corrected for between-subject or between-word variance).

RTs and the transformed RTs led to the same statistical conclusions for the primary questions. There were significant effects of frequency, priming, and contrast 1 (Exp 1 vs. Exp 2) as well as a significant interaction between priming and visual familiarity. Most important, the three-factor interaction of priming, frequency, and contrast 2 was significant when the log RT or reciprocal RT transformations were used, indicating a difference in the priming-frequency interaction effect seen in the Bodner and Masson study versus the Kinoshita experiment. We note, however, that this three-way interaction was not significant in the untransformed RT data. Separate LMMs for

TABLE 1
LMM estimates of fixed effects for untransformed RT, log RT, and reciprocal RT

	Measure								
	Untransformed RT			*Logarithmic RT*			*Reciprocal RT(-1/RT)*		
Effect	*Estim*	*SE*	*t*	*Estim*	*SE*	*t*	*Estim*	*SE*	*t*
(Intercept)	0.682	0.012	58.6	6.469	0.015	427.1	-1.624	0.023	-70.9
p	0.053	0.005	10.5	0.078	0.006	12.6	0.129	0.010	13.0
f	0.141	0.012	11.7	0.184	0.014	13.0	0.262	0.020	12.9
e.BM1-2	0.038	0.013	2.9	0.045	0.017	2.6	0.059	0.026	2.3
e.BM-SK	-0.071	0.008	-8.7	-0.103	0.011	-9.8	-0.166	0.016	-10.4
p:f	0.020	0.010	2.1	0.022	0.011	2.1	0.024	0.015	1.6
p:e.BM1-2	0.021	0.006	3.3	0.020	0.008	2.7	0.023	0.012	1.9
p:e.BM-SK	-0.005	0.004	-1.5	0.002	0.004	0.5	0.018	0.007	2.7
f:e.BM1-2	-0.002	0.009	-0.2	-0.011	0.010	-1.1	-0.024	0.014	-1.8
f:e.BM-SK	-0.051	0.008	-6.4	-0.060	0.009	-6.4	-0.074	0.013	-5.6
p:f:e.BM1-2	-0.001	0.012	-0.1	-0.006	0.013	-0.5	-0.012	0.019	-0.6
p:f:e.BM-SK	0.004	0.007	0.6	0.015	0.007	2.0	0.035	0.011	3.3

Note. p: priming, f: frequency, e.BM1-2: visual familiarity, e.BM-SK: semantic familiarity; ":" is a crossing operator.

the three experiments confirmed that the priming by frequency interaction was significant only in Experiment 3. Table 1 also shows that two of the lower order interactions of this three-factor interaction were significant for untransformed RT and log RT, but not for reciprocal RT. We also carried out the three corresponding rmANOVAs using subjects as random factor and obtained the same pattern of significant and non-significant effects regarding the three-factor interaction. Thus, as far as fixed effects are concerned, LMMs and rmANOVAs led to the same conclusions.

Variance/covariance component parameters. The variance/covariance component parameters for untransformed, logarithmic, and reciprocal RTs are listed in Table 2. They comprise the estimated standard deviations (i.e., square roots of variance estimates) of words' and subjects' means and of subject-related effects of frequency and priming as well as the associated estimates of correlations. The decisive role of the choice of transformation is apparent when we compare the estimates of correlations for the three RT variants. The correlation between priming and frequency effects is always estimated as positive, but there is a dramatic change of estimates

TABLE 2
LMM estimates of variance/covariance component parameters for untransformed RT, log RT, and reciprocal RT

	Measure								
	Untransformed RT			*Logarithmic RT*			*Reciprocal RT(-1/RT)*		
Effect	*SD*	*mean*	*priming*	*SD*	*mean*	*priming*	*SD*	*mean*	*priming*
Words									
mean	0.061			0.076			0.111		

Subjects									
mean	0.088			0.117			0.178		
priming	0.014	+0.392		0.027	-0.153		0.055	-0.482	
frequency	0.051	+0.557	+0.542	0.053	+0.137	+0.321	0.071	-0.342	+0.359
Residual	0.190			0.210			0.299		

Note. SD = square root of *lmer* variance estimate; remaining entries are estimates of correlations between effects.

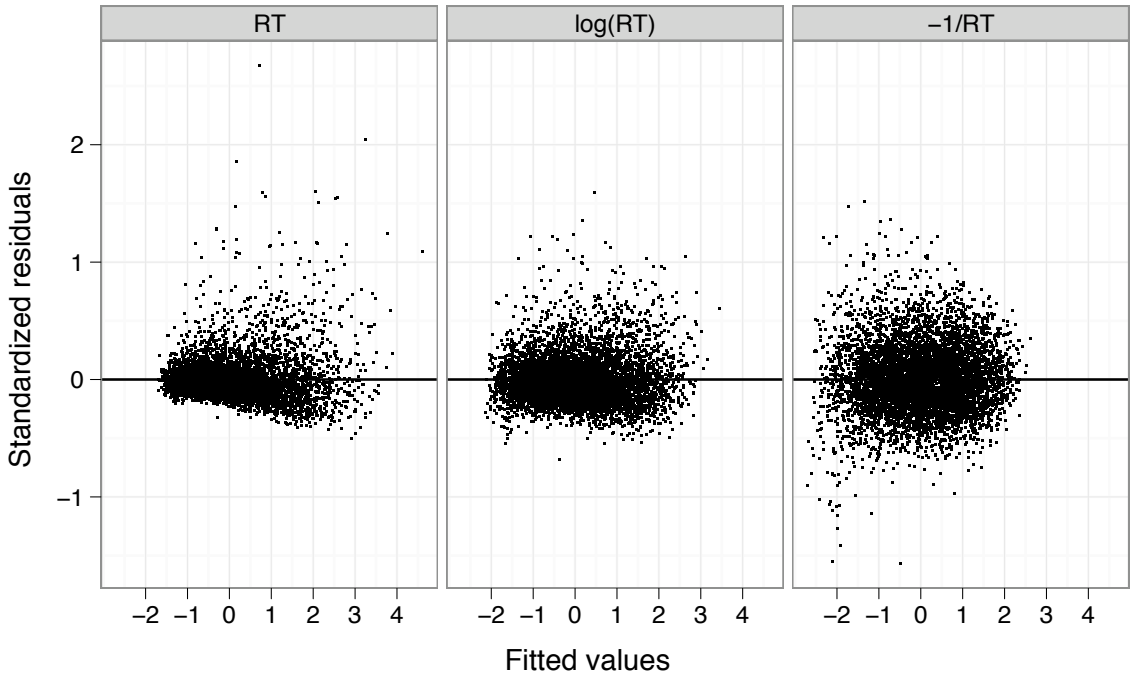

Figure 2. Plots of LMM residuals over normalized fitted values for untransformed (left), log transformed (middle), and reciprocal (right) RTs. Residuals in the bottom panel best meet model assumption of normality.

for the correlations of priming and frequency effects with mean RT when going from untransformed, to log-transformed to reciprocal RTs. Correlation estimates are highly *positive* for untransformed RTs, drop considerably in magnitude for log-transformed RTs, and end up as highly *negative* for reciprocal RTs (i.e., -1/RT).

Tests of correlation estimates. We tested the significance of correlation estimates by contrasting the LMM with an alternative LMM that assumes them to be zero. As the alternative is nested under the current model, we can compute a restricted likelihood-ratio based chi-square statistic. Forcing correlation estimates to zero led to a significant drop in goodness of fit for untransformed RTs, χ^2 (3) = 15.9, p = 0.001, and for reciprocal RTs, χ^2 (3) = 12.2, p = 0.007. Thus, the positive correlation estimates for untransformed RTs and the negative correlations for reciprocal RTs were significantly different from zero and, by implication, from each other. For log-transformed RTs, the corresponding increment in goodness of fit was not significant, χ^2 (3) = 2.4, p = 0.492. Thus, LMMs for standard RT and for reciprocal RT revealed significant correlation estimates of opposite sign between mean RT and priming and frequency effects.

Choice of transformation. The divergence of results for correlation estimates necessitates a decision for one of the three versions of RT. Inspection of standardized residuals plotted over fitted values suggests that standardized residuals for reciprocal RTs are closer to being normally distributed than those of the other two RT variants (see Figure 2). One quantitative method to decide on a suitable transformation is to estimate the optimal value of the λ–coefficient for the Box-Cox power transformation, $y(\lambda) = (y^{\lambda} - 1)/\lambda$, if $\lambda \neq 0$ and $y(\lambda) = \log(y)$, if $\lambda = 0$ (Box & Cox, 1964). The profile likelihood function for a range of values of λ can be determined with the *boxcox* function of the *MASS* package in *R* (Venables & Ripley, 2002), including also a horizontal line indicating an approximate 95%

likelihood-ratio confidence interval for the optimal value of λ. As shown in Figure 3, the optimal value of λ is -1.01 for subject-based mean RTs of experimental design cells. This value is very close to the value of -1 used for the reciprocal transformation of RT, indicating that this indeed is a very suitable simple transformation for these RT data. In contrast, the log transformation or keeping the RTs in the original metric would have been indicated for a $\lambda \approx 0$ or $\lambda \approx 1$, respectively. Thus, the Box-Cox transformation strongly suggests that reciprocal RTs are in a metric compatible with the normal-distribution assumption of our inferential statistics.

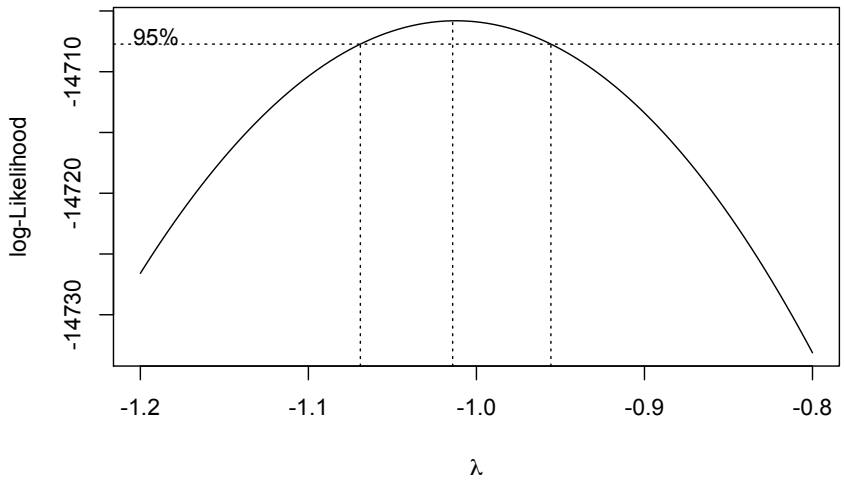

Figure 3. Profile likelihood function for λ, $L(\lambda)$ = const - $n/2\log(RSS(z^{(\lambda)})$, where $z^{(\lambda)} = y^{(\lambda)}/y'^{\lambda-1}$, y' is the geometric mean of the RTs, and $RSS(z^{(\lambda)})$ is the residual sum of squares for the regression of $z^{(\lambda)}$ (Box & Cox, 1964, cited after Venables & Ripley, 2002, p. 170 f.). The maximum of $L(\lambda)$ = 14705.81 at λ = -1.01. The graph also displays an approximate 95% likelihood ratio confidence interval for λ.

Discussion

The correlation estimates obtained for reciprocal RTs (i.e., -1/RT) indicate that subjects who respond faster exhibit stronger effects of the experimental manipulations. One interpretation of this relationship is that subjects who are more skilled readers, in the sense of being more efficient at encoding words, are more sensitive to differences in word frequency and to the influence of repetition primes. This latter possibility has potential implications for how researchers interpret masked priming effects. Such effects typically are believed to emerge without subjects being aware of the identity of the prime words or even that prime words are being presented (e.g., Forster & Davis, 1984). Finding larger priming effects for subjects who are better able to encode words may mean that those subjects are more likely to be aware of the masked primes. Awareness of the primes could enhance the size of the priming effects and this possibility suggests the need for more rigorous testing of the assumption that masked priming occurs without awareness. Alternatively, if we interpret response speed as an indicator of task engagement or adherence to instruction, the result suggests that subjects who engage more will be more likely to reveal the experimental effects. Thus, in situations of low statistical power, investment in motivational incentives may pay off.

In addition, the sign of the correlation estimate between mean response speed and effect size is in clear contradiction to the assumption of a general speed factor (i.e., subjects who are slow in general are also the ones who show larger experimental effects). As we pointed out, both hypotheses have guided interpretations of RT-difference based correlations. Our results provide a clear answer for this question for masked-repetition priming and frequency effects in the lexical decision task.

There is also a positive correlation estimate between priming and frequency effects. This correlation estimate hints at a possible common mechanism shared by processes that generate repetition priming and processes that are sensitive to word frequency. Of course, with only two effects measures we are not in a position to infer anything about the specificity and generalizability of these effects. Experimental designs with a larger number of factors, however, may inform about dissociations between experimental effects via differential correlation estimates.

CORRELATIONS

The analyses in the previous section established that reciprocal transformation of RT leads to a specific pattern of correlation estimates between mean RT and experimental priming and frequency effects that is different from that obtained from untransformed RTs. In a LMM, these correlations are estimated as parameters simultaneously with the fixed effects. How do these correlation estimates compare with correlations computed from scores based on the individual subjects' data, that is, from within-subject analyses obtained in a repeated-measures multiple regression

TABLE 3
Standard deviations (SD) and correlations of reciprocal mean (-1/RT), frequency effect, and priming effect based on within-subject analysis

	SD	*mean*	
	priming		
mean	0.297		
priming	0.089	-0.365	
frequency	0.138	+0.439	-0.177

Note. Compare the within-subject correlations with corresponding LMM estimates of correlations in Table 2.

analysis (rmMRA; Kliegl et al., 2006; Lorch & Myers, 1990)? The correlations based on these within-subject reciprocal RTs are presented in Table 3 and can be compared directly with corresponding LMM estimates in Table 2.

The divergence between correlations based on within-subject reciprocal RTs and LMM model estimates is substantial. In particular, the two correlations with the frequency effect are opposite in sign. They are a consequence of the notorious unreliability of within-subject difference scores. Interestingly, problems of reliability may not only reduce the magnitude of correlations (e.g., from +0.40 to -0.18 for the correlation of frequency and priming effects), but they may also suggest a correlation of similar magnitude though of opposite sign, relative to that generated by a more reliable method (e.g., +0.44 instead of -0.34 in the case of the correlation between mean RT and frequency effect). The LMM corrects for the unreliability of individual subjects' scores by "borrowing strength" from the presumably reliable population estimate afforded by the complete sample. Essentially, (a) the more extreme an observed mean, (b) the smaller the number of observations, and (c) the larger the variance for a given subject's data, the more will this subject's conditional mean be based on the overall mean (i.e., "shrunken" towards the population mean; see Gelman & Hill, 2007, especially chapters 12 and 18, for expositions).[3]

The LMM model estimates listed in Tables 1 and 2 can be used to generate predictions for each subject's mean RT as well as each subject's priming and frequency effects--the so-called best linear unbiased predictions (BLUPs; Henderson, 1953), more appropriately referred to as the conditional means evaluated at the estimated parameters (Bates, 2008a). In other words, the LMM-based adjustments of means and effects are calculated *after* estimation of the variance/covariance component parameters. Thus, formally, these adjustments are not parameters of the model. Figure 4 displays the 95% prediction intervals for conditional means and conditional priming and frequency effects for 72 subjects, sorted by the conditional means (-1/RT) and centered on corresponding fixed effects. Subjects' prediction intervals overlap more strongly for priming and frequency effects (middle and right panels) than for mean RTs (left panel), because of the lower reliability of difference scores. Large mean (-1/RT) values tend to go along with small priming and frequency effects, in agreement with the LMM estimates of correlations (see Table 2).

The scatterplots of filled circles in the panels of Figure 5 represent the conditional means of 72 subjects predicted from the LMM estimates in the more familiar format. In addition, we also plot the unadjusted within-subject effects as open circles and arrows pointing from within-subject values to their corresponding conditional

[3]Gelman and Hill (2007) illustrate shrinkage for the case of a model without predictors. Applied to our data, if M is the overall mean RT, m_j and n_j are mean and number of RTs of subject j, σ^2_y and σ^2_α are residual and between-subject variances, then the predicted mean RT α_j for subject j can be approximated as a weighted average of the subject's mean RT and the overall mean RT:

$$\alpha_j \approx [(n_j / \sigma^2_y)\, m_j + (1/\sigma^2_\alpha)\, M] / [(n_j/\sigma^2_y) + (1/\sigma^2_\alpha)]$$

Then, for the limiting case of $n_j = 0$, $\alpha_j = M$, and for $n_j \rightarrow \infty$, $\alpha_j = m_j$. Thus, on the one hand, the fewer RTs contributed by a subject, the stronger is the overall mean's contribution to the predicted mean for this subject; indeed, in the case of missing data

($n_j = 0$), we simply predict the overall mean M. On the other hand, the larger the number of RTs, the more the prediction is based on the observed subject's mean. Weights also depend on the ratio of residual and between-subject variances. For example, for $n_j = \sigma^2_y/\sigma^2_\alpha$, subject and overall mean are equally weighted in the prediction; that is, the formula reduces to $\alpha_j = ½ (m_j + M)$. Assuming constant residual variances for subjects (which is not necessary in general), if $n_j > \sigma^2_y / \sigma^2_\alpha$ (i.e., for large differences between subjects relative to the number of observations for subject j and the residual variance), α_j will move towards m_j; conversely, if $n_j < \sigma^2_y / \sigma^2_\alpha$ (if there is large residual variance or if there are few observations), α_j will move towards M.

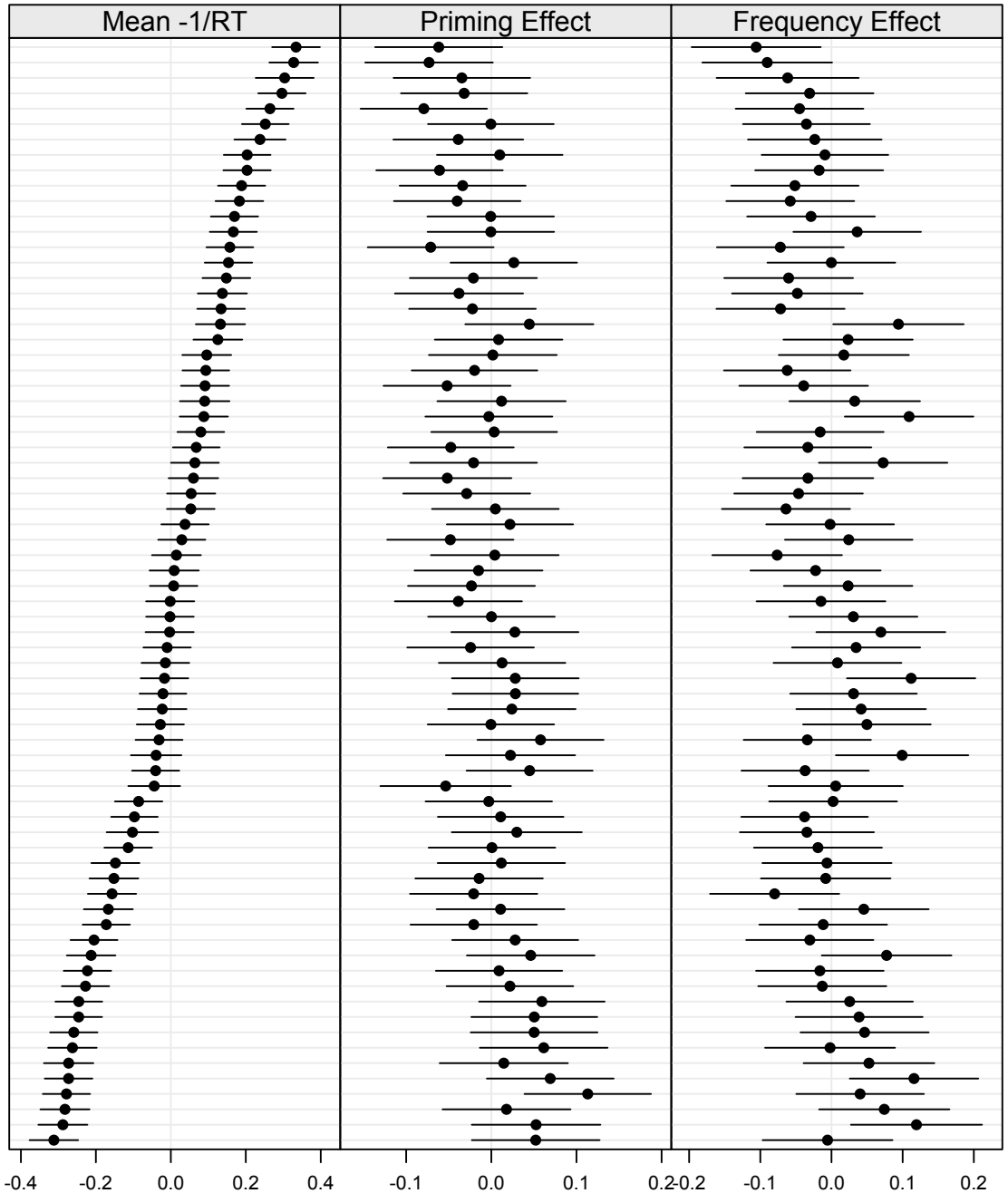

Figure 4. Caterpillar plots for 72 subjects' (a) conditional means (-1/RT), (b) priming effects, and (c) frequency effects, centered on corresponding fixed effects. Subjects are ordered by mean (-1/RT). Horizontal lines indicate 95% prediction individuals; R script adapted from Bates (2008c).

means. The fact that the arrows point from the outside towards the center of the plot illustrates the model-based shrinkage due to unreliability of within-subject values. In general, the more extreme a value, the longer is the arrow, implying more shrinkage towards the population mean for such values. As a consequence of this shrinkage,

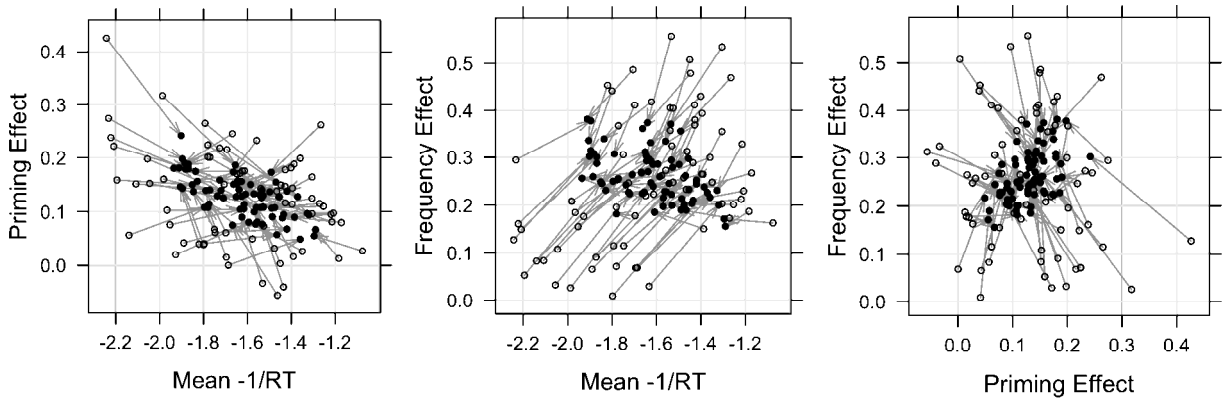

Figure 5. Scatterplot of within-subject means (open symbols) and conditional means (filled symbols). Arrows connect the two values for each subject. Shrinkage correction changes correlations between mean reciprocal (-1/RT) and priming effect (left), mean reciprocal (-1/RT) and frequency effect (middle), and between priming and frequency effects (right); R script adapted from Bates (2008c).

however, the "correlations" of conditional means may take on a different sign than the corresponding within-subject correlations. Most noticeable in the present data, the correlation between reciprocal mean (-1/RT) and priming effect increases (left panel), the positive within-subject correlation changes from a positive to a negative correlation for reciprocal mean (-1/RT) and frequency effect (middle panel in Figure 5), and the correlation between frequency and priming effect changes from a low negative correlation to a medium positive one.

Once a correlation structure is unveiled as nicely as in Figure 5, one is tempted to carry on with the conditional-mean predictions based on the LMM parameter estimates (i.e., the data represented by filled circles) to compute, for example, correlations with other background variables such as IQ or age that may be available for each subject. Unfortunately, however, this would be an erroneous step and would take the LMM-based predictions too far. First, given that conditional means represent a compromise between the subjects' means and the estimate of the population mean, they must obviously not be treated as independent observations. Second, "correlations" based on conditional means (i.e., the filled circles in the panels of Figure 5) are not identical to the LMM correlation estimates. Actually, "correlations" of conditional means tend to be larger in absolute magnitude than the corresponding LMM correlation parameters (see next section). Therefore, subject-level variables such as IQ and age effects must be incorporated as covariates in the LMM where, of course, one can also specify and test interactions between effects of individual differences (e.g., age, IQ) and experimental effects (e.g., priming or frequency manipulations).

SIMULATION OF LMM ESTIMATES OF VARYING INTERCEPTS AND SLOPES

Conditional means have much appeal because they take into account differences in the reliability of the grouped data. As a stern warning against treating them as independent observations, however, we document with simulations that it is indeed the LMM estimates, not the conditional means predicted from them, that we have to rely on for inference. To this end, we generated 100,000 sets of data for a simple LMM model including 30 "subjects" and a predictor with 10 levels, conforming to a known variance for intercept and slope across subjects and varying the true correlation between these parameters from -0.9 to +0.9 in 2,000 steps (i.e., each simulation used a different correlation). Results are displayed in Figure 6. The x-axis in each panel represents the true correlation in the simulation. The panels in the top row, from left to right, plot the difference between model estimates and true values for (a) intercept variance, (b) slope variance, (c) their covariance as well as (d) the derived correlation. The dashed horizontal line is the reference line for perfect recovery of the true values. The black line lists the mean estimate for all data and the grey line lists the mean estimate when 50% of a data set was randomly deleted before model estimations (i.e., an extreme form of model estimation with missing data). In general, model estimates are close to the zero line, indicating that the parameters were recovered quite well.

The second row plots the differences between conditional means and true values. It is immediately apparent that

conditional means underestimate variances and exaggerate covariances and correlations. The shrinkage of variance reflects the contribution of the likelihood in the computation of conditional means. Shrinkage correction for predictions leads to dampening of the variance components, but, as we have shown in this section, not of the associated covariance component. The shrinkage of variance prevents overfitting of unreliable data but, as a curious

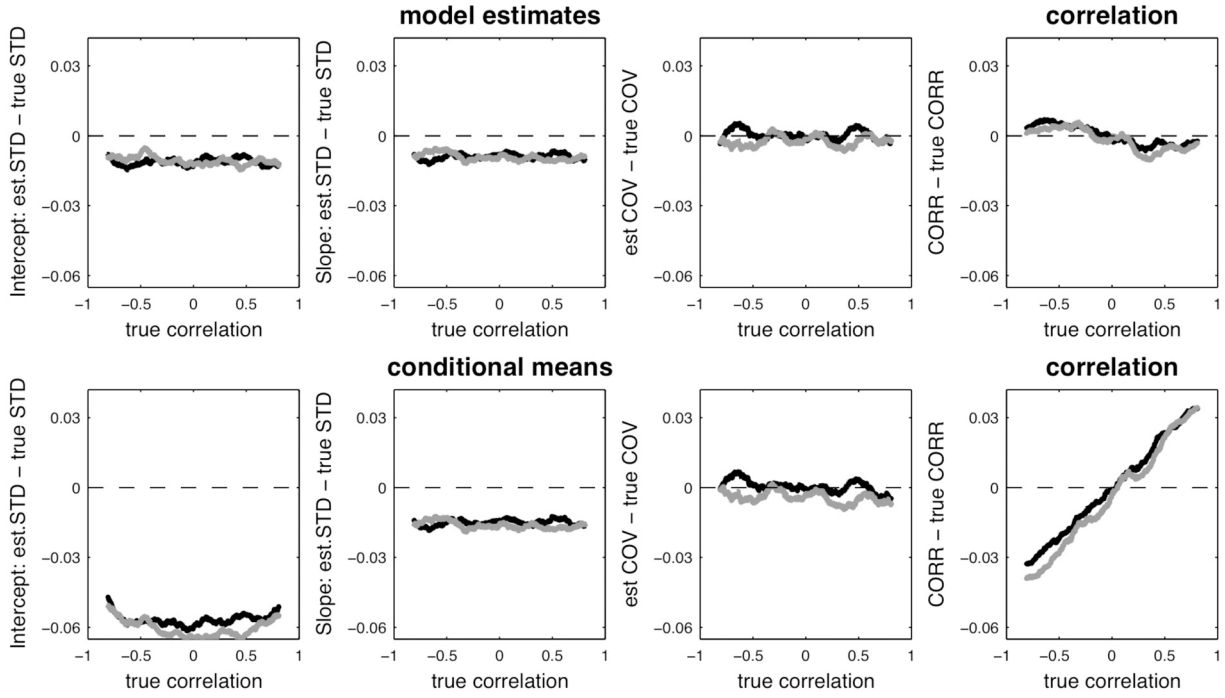

Figure 6. Results of 100,000 LMM simulations with varying slopes and intercepts. Standard deviations for variance-component parameters were fixed and correlations for each run were known. The plots show moving averages of LMM estimates (top) and conditional means (bottom) as functions of respective true correlation; the columns contain estimates of standard deviations and covariance, and of the respective correlation. Black and gray lines denote runs with balanced and unbalanced (i.e., random 50% of data) designs.

side effect, the "correlations" based on conditional means for individual subjects are larger in absolute value than the corresponding LMM estimates of the correlation.

In summary, the simulation shows that model estimates (top) are fine, even when estimates are based on only 50% of the data, but conditional means (bottom), because of their dependency on estimated population values, clearly overestimate the magnitude of the correlation estimate. Thus, interpretations of correlations must not be based on the "correlations" computed from conditional means, but on LMM estimates of correlations.

GENERAL DISCUSSION

Analyzing RT Experiments with LMM

We re-analyzed data from fairly typical RT-based experiments that examined effects of masked repetition priming, word frequency, visual familiarity, and semantic familiarity as well as their interactions. The goal of this article was to demonstrate that, in addition to these fixed effects of experimental manipulations, we can also simultaneously estimate the variances and correlations of mean RT, priming effects, and frequency effects for subjects in a LMM. RTs enter the analysis in an unaggregated format, rather than being averaged over words within design cells prior to analysis in a mixed-model ANOVA with subjects as a random factor. The new analysis requires the specification of subjects and words as crossed random factors and assumes that means and effect sizes are normally distributed among subjects and among words. The term "random" is a bit unfortunate in this context; it derives from the (presumably) random sampling of subjects and words; it definitely does not refer to anything that could be dismissed. There is much to be said for Gelman and Hill's (2007) proposal to refer to these variances as based on varying intercepts and varying effects.

The validity of variance/covariance component parameters, in particular the correlations between mean RT and frequeny and priming effects among subjects, depended on a reciprocal transformation of RTs which was required to establish normality for residuals of fitted models. Given a suitable transformation of RTs, we established fixed effects of priming, frequency, visual familiarity, and semantic familiarity as well as interactions among them. The results were in agreement with those of the original publications; in particular, the interaction between frequency and priming was not significant for the Bodner and Masson (1997) data, but was significant for the Kinoshita (2006) data. Only the use of transformed data, however, revealed a statistically significant difference across experiments with respect to the frequency by priming interaction.

Novel results were obtained when the reciprocal transformation was used: (1) negative correlation estimates between mean (-1/RT) and both priming and frequency effects, and (2) a positive correlation estimate between frequency and priming effects among subjects. These results indicate that we observe stronger experimental effects for faster subjects. Such a pattern could emerge if short RTs indicate a difference in encoding efficiency. Alternatively, short RTs may reflect task engagement, with strong task engagement leading to stronger experimental effects. Obviously, with only an overall mean and two effects, we are not in a position to prefer one of these explanations to the other. The results do rule out, however, the proposition that longer RTs are a prerequisite for observing stronger experimental effects. Both LMM estimates of correlations and within-subject correlations based on untransformed RTs would have led us to erroneously support this claim.

Correlations based on within-subject analysis are generally weaker than the correlations estimated in a LMM. Much research on individual differences in cognitive processes has been invested in correlating difference scores related to experimental conditions, only to find that the inherent unreliability of these scores rendered them close to useless for establishing a link to traditional psychometric research. For example, demonstrations of low reliability in priming effects, such as those reported by Stolz et al. (2005), may have been influenced by this problem. From a LMM perspective, researchers using the typical difference score approach have stacked their cards heavily against themselves for finding reliable relationships. This problem generalizes to the use of repeated-measures multiple regression analysis (e.g., Kliegl et al., 2006; Lorch & Myers, 1990), which has also been shown to run the risk of being anti-conservative (Baayen et al., 2008). Nevertheless, sufficiently large sample sizes will remain a prerequisite for estimating correlations between experimental effects even with a LMM. In the present report, for example, significant correlation estimates depended on aggregating over three experiments. Moreover, in this case an appropriate transformation of the raw data was required as well to ensure the validity of the analysis.

Theoretical Relevance for Priming Studies

This set of data and results is not unique. Indeed, alternating case has been shown to exaggerate word-frequency effects in the word-naming task (Besner & McCann, 1987; Herdman et al., 1999). The effect of introducing some form of degradation of the visual form of words, however, does not produce a consistent influence on performance. Visual degradation achieved through contrast reduction has been shown to be additive with word frequency (Borowsky & Besner, 1991, 1993), suggesting that frequency-sensitive processes are rather late in the chain of processing components leading to word recognition--at least late enough to miss all the action of early perceptual processes that establish a clear signal against variable noise background. Interestingly, in the same experiments that show additivity between contrast and frequency, there is clear evidence for much larger semantic priming effects under conditions of low visual quality, in line with the argument that slowing causes an increase of effects riding on RT. Our LMM analyses show, however, that at the level of individual differences, it is the faster subjects who show larger priming and word-frequency effects. It is an open and theoretically interesting question as to whether a similar pattern would emerge if contrast were used as the means of visual degradation and semantic rather than masked repetition priming were used as the method of manipulating context. Indeed, it is quite possible that faster subjects will show less priming under conditions that present primes in a clear and easily seen format, but targets in degraded form.

To Transform or not to Transform?

Invariably, questions are raised about the justification for choosing a transformation of the dependent variable. This issue has been addressed many times in the psychological research literature with not much impact. An exception is that, compared to 20 years ago, it seems that logarithmic transformation of RTs needs no justification. One reason probably is that, for the assessment of fixed effects, it rarely matters. Our analyses support the assumption that this kind of tacit knowledge guides individual decisions about the transformation question. Irrespective of whether we look at untransformed, log-transformed, or reciprocal RTs, we consistently obtain significant main effects of priming, frequency, and visual familiarity. Matters change substantially, however, when we look at the correlations

of individual differences or estimates of such correlations between these effects. Transformation may not influence the pattern of differences between means, but it can drastically alter the pattern of correlations between effects or between effects and mean response speed.

So which metric is the correct one? Sometimes psychologists do not use transformations such as those suggested by the Box-Cox procedure because they perceive RTs as the natural metric. In the linear model, coefficients reflect the additional time due to an experimental effect; that is, the time it takes for a hypothetical cognitive process to finish. Thus, they give priority to additivity of time and attempt to explain the general positive skew of RT distributions and their heteroscedasticity across conditions as a consequence of internal information processing (e.g., Logan, 1992; Wagenmakers & Brown, 2007).

In contrast, effects estimated in a transformed RT metric may not have an obvious interpretation. There is some force to this argument, but at least the two transformations considered in this article do have psychologically plausible interpretations: 1/RT leads to an interpretation of coefficients as additive changes in processing rate possibly tied to neural spike rates (e.g., Carpenter, 1981; Carpenter & Williams, 1995) and coefficients estimated from log(RT) inform about the size and reliability of standard RT effects in multiplicative, rather than additive, terms. Thus, we can also develop models with these metrics.

The general problem, however, is that, if the linear model is to be used for statistical inference, then it simply does not make sense to work with a yardstick for which the precision of measurement changes with the size of the object to be measured. The only generally applicable (i.e., independent of specific content domains) and meaningful estimate of the precision of our measurement scale is the standard deviation. Therefore, if statistical inference is intended for fixed and random experimental effects, one solution is to transform one's scale such that the same standard deviation holds across the entire range, a characteristic that often does not hold when untransformed RT data are considered (Wagenmakers & Brown, 2007). The reciprocal transformation appears to achieve this goal for the RT data considered here; for other data sets, logarithmic, square-root, or no transformations may be called for.

Nevertheless some theoretical constructs make perfect sense in one metric, but not in another. So, can't we have our cake and eat it, too? The standard linear model requires a normally distributed measure, but RTs obviously do not have this property. They appear, however, to be well described, for example, by lognormal or gamma distributions. If one is theoretically committed to such a distribution (e.g., the SWIFT model of eye movement control in reading randomly samples the starting times of saccade programs from a gamma distribution; Engbert, Nuthmann, Richter, & Kliegl, 2005), then an elegant solution, one that preserves interpretation in the standard RT metric, is to switch from the linear mixed model to a generalized linear mixed model (GLMM) for statistical inference. The disadvantage associated with this approach is that estimation of GLMM coefficients (in particular from crossed-random effects models) must be numerically approximated rather than computed from a closed-form solution (Bates, 2008b). Also the interpretation of coefficients is less straightforward for GLMM than for LMM.

There are also recent developments to estimate crossed-random effects of subjects and items in a Bayesian framework (Rouder, Lu, Speckman, Sun, & Jiang, 2005). This approach opens the way to use distributions outside the exponential family. It also allows other than a normal parent distribution for the parameter estimates. For example, Rouder et al. showed that RTs for symbolic distance effects (e.g., judging the difference between numerically adjacent and non-adjacent digits) are best described with a three-parameter Weibull distribution, assuming also that the parameters themselves are gamma distributed. Their Bayesian estimation of a hierarchical Weibull model also represents an alternative to maximum likelihood estimation; simulations prove the Bayesian approach to be superior to eight alternative estimation methods for Weibull parameters at the individual or group level. The advantage becomes especially striking when simulations are based on only 20 rather than 80 observations per subject. Finally, Rouder et al. point out that their approach can be expanded to achieve what has been described in this article: The simultaneous estimation of variance/covariance component parameters for subject and items.

In conclusion, the routine application of Bayesian techniques will still take a few years to take hold in experimental psychology. In the mean time, we propose to spend one degree of freedom for a transformation that maps the observed data into a representation that is compatible with the statistical model we use for inferences or prediction or to use GLMM to this end. Typically, either approach will yield normally distributed residuals. In this measurement space, we interpret not only the fixed effects of our experimental design but, in a single sweep, we can estimate how these effects vary and correlate among subjects and items. Experimental psychologists have collected much reliable, theoretically relevant information on subjects and items for many years. Perhaps the time has come to use it.

REFERENCES

Baayen, R. H. (2008) *Practical data analysis for the language sciences with R*. Cambridge, MA: Cambridge University Press.

Baayen, R.H., Davidson, D.J., & Bates, D.M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59,* 390-412.

Balota, D. A., Yap, M. J., Cortese, M. J., & Watson, J. M. (2008). Beyond mean response latency: Response time distributional analyses of semantic priming. *Journal of Memory and Language*, *59,* 495-523.

Bates, D. M., Maechler, M., & Dai, B. (2009). *lme4: Linear mixed-effect models using S4 classes*. R package version 0.999375-28. [Software]. Vienna, Austria: R Foundation for Statistical Computing.

Bates, D. M. (2008a). *Linear mixed model implementation in lme4.* Technical Report.

Bates, D.M. (2008b). *Computational methods for mixed models.* Technical Report.

Bates, D.M. (2008c). *Fitting linear mixed-effects models using the lme4 package in R.* Presentation at Potsdam GLMM workshop, August 7, 2008.

Becker, C. A., & Killion, T. H. (1977). Interaction of visual and cognitive effects in word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 3*, 389-401.

Besner, D., & McCann, R. S. (1987). Word frequency and pattern distortion in visual word identification and production: An examination of four classes of models. In M. Coltheart (Ed.), *Attention and performance XII: The psychology of reading* (pp. 201-219). Hillsdale, NJ: Erlbaum.

Birren, J.E. (1956). The significance of age changes in speed of perception and psychomotor skill. In J.E. Anderson (Ed.), *Psychological aspects of aging.* Washington, DC: American Psychological Association.

Bodner, G. E., & Masson, M. E. J. (1997). Masked repetition priming of words and nonwords: Evidence for a nonlexical basis for priming. *Journal of Memory and Language, 37,* 268-293.

Borowsky, R., & Besner, D. (1991). Visual word recognition across orthographies: On the interaction between context and degradation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*, 272-276.

Borowsky, R., & Besner, D. (1993). Visual word recognition: A multistage activation model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 813-840.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society B, 26,* 211-252.

Carpenter R. H. S. (1981). Oculomotor procrastination. In D. F. Fisher, R. A. Monty, & J. W. Senders (eds.), *Eye movements: cognition and visual perception* (pp 237-246). Hillsdale, NJ: Erlbaum.

Carpenter , R. H. S., & Williams, M. L. L. (1995) Neural computation of log likelihood in the control of saccadic eye movement. *Nature, 377,* 59-62.

Coltheart, M. (1981). The MRC Psycholinguistic database. *Quarterly Journal of Experimental Psychology, 33A,* 497-508.

Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review, 112,* 777-813.

Faraway, J. J. (2006). *Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models.* Boca Raton, FL: Chapman & Hall/CRC.

Forster, K. I., & Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 680-698.

Forster, K. I., Mohan, K., & Hector, J. (2003). The mechanics of masked priming. In S. Kinoshita & S. J. Lupker (Eds.), *Masked priming: The state of the art* (pp. 3-37). New York: Psychology Press.

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

Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics, 9, 226-252.*

Herdman, C. M., Chernecki, D., & Norris, D. (1999). Naming cAsE aLtErNaTeD words. *Memory & Cognition, 27*, 254-266.

Kinoshita, S. (2006). Additive and interactive effects of word frequency and masked repetition in the lexical decision task. *Psychonomic Bulletin & Review, 13,* 668-673.

Kliegl, R. (2007). Toward a perceptual-span theory of distributed processing in reading: A reply to Rayner, Pollatsek, Drieghe, Slattery, & Reichle (2007). *Journal of Experimental Psychology: General, 136*, 530-537.

Kliegl, R., Nuthmann, A., & Engbert, R. (2006). Tracking the mind during reading: The influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology: General, 135,* 12-35.

Loftus, G. R. (2002). Analysis, interpretation, and visual presentation of experimental data. In H. Pashler (Ed.), *Stevens' handbook of experimental psychology* (Vol. 4, pp. 339-390). New York: John Wiley and Sons.

Logan, G. D. (1992). Shapes of reaction-time distributions and shapes of learning curves: A test of the instance theory of automaticity. *Journal of Experimental Psychology: Language, Memory and Cognition, 18,* 883-914.

Lorch, R. F., & Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Language, Memory and Cognition, 16,* 149-157.

R Development Core Team (2009). *R: A language and environment for statistical computing*. (version 2.8.1). [Software]. Vienna, Austria: R Foundation for Statistical Computing.

Ridderinkoff, K. R. (2002). Micro- and macro-adjustments of task set: Activation and suppression in conflict tasks. *Psychological Research, 66*, 312-323.

Rouder, J. N., Lun, J., Speckman, P., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review, 12,* 195-223.

Sarkar, D. (2008). *Lattice: Multivariate data visualization with R*. New York: Springer. [Software] R package version 0.17-20.

Steinhauser, M., & Hübner, R. (2008). How task errors affect subsequent behavior: Evidence from distributional analyses of task-switching effects. *Memory & Cognition, 36*, 979-990.

Stolz, J. A., Besner, D., & Carr, T. H. (2005). Implications of measures of reliability for theories of priming: Activity in semantic memory is inherently noisy and uncoordinated. *Visual Cognition, 12*, 284-336.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S.* New York: Springer. [Software] R package version

7.2-45.
Wagenmakers, E.-J., & Brown, S. (2007). On the linear relation between the mean and the standard deviation of a response time distribution. *Psychological Review, 114*, 830-841.
Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & McKoon, G.(2008). A diffusion model account of criterion shifts in the lexical decision task. Journal of Memory and Language, 58, 140-159.
Wickham, H. (2007). Reshaping data with the *reshape* package. *Journal of Statistical Software, 21,* 1-19. [Software] R package version 0.8.2.
Wickham, H. (2009). *ggplot2*: elegant graphics for data analysis. New York: Springer. [Software] R package version 0.8.1 http://had.co.nz/ggplot2/