PSIHOLOGIJA, 2018, Vol. 51(4), 469–488
© 2018 by authors

UDC 159.953.3.075
DOI: https://doi.org/10.2298/PSI170615023P

Analyzing data from memory tasks – comparison of ANOVA, logistic regression and mixed logit model

Milica Popović Stijačić[1**], Ljiljana Mihić[2**], and
Dušica Filipović Đurđević[1,2**]

[1] Laboratory for Experimental Psychology,
University of Novi Sad, Serbia
[2] Department of Psychology, Faculty of Philosophy,
University of Novi Sad, Serbia

We compared three statistical analyses over binary outcomes. As applying ANOVA over proportions violates at least two classical assumptions of linear models, two alternatives are described: the binary logistic regression and the mixed logit model. Firstly, we compared the effects obtained by the three methods over the same data from a previous memory research. All three methods gave similar results: the effects of the tasks and the number of sensory modalities were observed, but not their interaction. Secondly, by using the bootstrap estimates of the parameters, the efficacy of each method was explored. As predicted, the bootstrap parameter estimates of the ANOVA had large bias and standard errors, and consequently wide confidence intervals. On the other hand, the bootstrap parameter estimates of the binary logistic regression and the mixed logit models were similar – both had low bias and standard errors and narrow confidence intervals.

Key words: cued/free recall, ANOVA, logistic regression, mixed logit model, bootstrapping

Corresponding author: milica.popovic657@gmail.com

Acknowledgement. This research is funded by Ministry of science and environmental protection of Republic of Serbia (grant numbers: 179033 and 179006). The results of this study were presented at “XXII naučni skup Empirijska istraživanja u psihologiji”, Belgrade, Serbia, March 18–20, 2016, and the authors would like to express their gratitude to the audience at the conference for helpful comments. The authors are also grateful to Nemanja Vaci, Petar Milin, Tanja Jevremov, and Milena Kresoja for helpful consultations.

[*] All authors equally contributed to the paper.

[1,2] Dušica Filipović Đurđević has moved and currently holds a position at the Department of Psychology, Faculty of Philosophy, University of Belgrade, Čika Ljubina 18–20, 11000 Beograd, Serbia, Tel: +381 11 2630 542, Fax: +381 11 2639 356, e-mail: dusica.djurdjevic@f.bg.ac.rs

Highlights:

- Three statistical analyses compared for binary outcomes in a memory task.
- Similar pattern of results with ANOVA, binary logistic regression and mixed logit model.
- Bootstrap analysis of the estimates revealed the fine-tuned differences.
- ANOVA was related to large bias and standard errors, and wide confidence intervals.
- Logistic regression and mixed logit models recommended for binary outcomes.

When analyzing data from memory tasks, researchers usually perform statistical methods from the family of general linear models (GLM; Jeager, 2008; Quene & van der Bergh, 2008). Basic linear model is often presented as $y = \beta 0 + ß1X + \mathcal{E}$, where y represents dependent or response variable, x is independent or predictor variable, $\mathcal{E}$ is a random term of the equation, or the error term. Using x and y observed values we than estimate $\beta 0$ and $\beta 1$ regression coefficients. The parameter estimation in linear modelling is usually done by least squares estimation. ANOVA models can be presented as the special case of linear models $yij = \mu + \alpha i + \mathcal{E}ij$, where μ is mean, αi is the main effect of an i-th level of the factor (Rencher & Schaalje, 2008). In linear models, dependent variable must be continuous; however, data from memory tasks are often categorical, or more precisely binomial. In practice, researchers calculate proportion of correct answers over stimuli and participants, and then, apply some of the GLM methods, with the proportion of correct responses as the dependent variable. Such practice is not legitimate given two assumptions of linear models. The first one is attributed to the violation of *homoscedasticity,* when the variance of the proportion is not equally distributed around the regression line. The reason for this is that the variance is not independent from its mean and exceeds the maximum for the mean proportion of $p = .50$. This conclusion follows directly from the equation for the variance of the population proportion: $\sigma^2 =$ (for more detailed explanation see Jaeger, 2008). The second violation is related to the assumption that a dependent variable can take any real value, which is in this case violated due to the proportion being bounded between 0 and 1 (Baayen, 2012). These violations of the linear model assumptions increase the risk of Type I and Type II errors, and consequently, decrease the power of statistical tests.

The solution for these problems is to apply the appropriate statistical methods from the family of generalized linear models (Agresti, 2002; Baayen, 2008; Ferrari & Comelli, 2016; Jaeger, 2008; Murayama, Sakaki, Yan, & Smith, 2014; Quené & van der Bergh, 2008), such as binary logistic regression and mixed logit model. In both of these methods, binary distributed outcomes (correct and incorrect answers) are transformed via logit transformation (natural logarithm of odds ratio):

$$\text{logit (Y)} = \text{log.}\left(\frac{\text{succes}}{\text{fails}}\right) = \text{log}(\frac{\text{p}}{1-\text{p}})$$

In this case, the dependent variable has a desirable range of values (which are bounded between minus infinity and plus infinity) and a linear relationship with the predictors:

$$\text{logit (Y)} = \beta 0 + \beta 1\text{X}1 + \beta 2\text{X}2 +...+ \beta i\text{X}i = \text{x`}\beta.$$

In the equation above, logit(Y) is the predicted value of logit(p), β_0 is an estimate of the intercept, and βi is the regression coefficient estimate of the predictor Xi (at the end, the right term of the equation x`β, is presented using vector notation). Estimations of the coefficients in the model are fitted in accordance with the maximum likelihood estimation method (Jaeger, 2008; Tabachnick & Fidell, 2007).

Mixed logit models describe binary dependent outcome as the linear combination of fixed and random effects (Jaeger, 2008). The fixed effects are written in the same way as in the binary logistic regression model as x`β (x` contains the data related to the predictors, and β contains the fixed effect coefficients). However, the important difference lays in the possibility of modelling random effects related to participants or/and stimuli z`*b* (z` are the data related to participants or stimuli, and *b* contains random effect coefficients). As per vector notation, the whole model is written as:

$$\text{logit}(p) = \text{x`}\beta + \text{z`}b.$$

Unlike the logistic regression models, the fit of the mixed logit model is reached through the *quasi-log-likelihood* – computational optimization of the approximation of the true log-likelihood (Jaeger, 2008).

This potential of the random effect modelling is highly important, especially for the application in memory research (Murayama et al., 2014). On the one hand, there are individual differences due to memory ability (Brewer & Unsworth, 2012). On the other hand, when using words as stimuli in any research, it is impossible to exhaust all potential lexical and linguistic variations. Consequently, it is justified to treat word stimuli as a random variable in the analysis (Baayen, Davidson, & Bates, 2008). Thus, mixed effects can serve as a solution to the well-known *language-as-fixed-effect-fallacy* (Baayen et al., 2008; Clark, 1973; Coleman, 1964; Quené & van der Bergh, 2008; Raaijmakers, Schrijnemakers, & Gremmen, 2008). This refers to the peculiarity of psycholinguistic research (but also any type of research that uses verbal stimuli) that the sampling is conducted both from the population of participants, and from the population of words. At the same time, the goal of the research is to generalize the findings to both populations – to expect to find the same observations on different participants and by using different words with the specified features. In order to treat participants and language items as random variables, researchers have traditionally performed two ANOVA analyses. Consequently, two *F* tests would be obtained, where *F1* test comes from the by-participant analysis, and *F2* test comes from the by-item analysis. Prior to the publication of the Clark`s article (Clark, 1973), researchers had reported only *F1* test in their papers. Clark influenced the scientific community to treat language as a random effect as well. In addition, a new unit of measurement was developed – the *quasi F test*, in order to simultaneously estimate the size of *F1* and *F2* tests. Unfortunately, this recommendation has not been applied as the aforementioned one, probably due to its complicated computation, and some other limitations (Raaijmakers et al., 2008). However, avoiding presenting the *quasi F test* is not correct, because it was shown by Raaijmakers et al. (2008) that reporting two statistically significant *F* tests was not enough. Even when *F1* and *F2* tests are statistically significant, *quasi F test* could remain beyond the critical *p* value, and consequently could increase the probability of Type II error. The mixed effect models solved this problem due to the fact that they could capture variations from both participants and language items. In addition to accounting for the general variability of participants and items, these models can capture a more fine grained variation in the slopes of the effects of interest (i.e., fixed effects) by allowing them to vary across participants and items. Capturing the variation from random effects consequently decreases variance of fixed effect estimates (Clark & Linzer, 2015; Gelman & Hill, 2007). In other words, this class of models is able to account for individual sensitivity to the effects that are under examination (e.g., different sensitivity to word frequency effect of participants with varying reading skills). Furthermore, introduction of the random effects has one additional benefit: it decreases variance of fixed effect estimates. The advantages of mixed effects models have been discussed within various scientific disciplines (Baayen et al., 2008; Bolker et al., 2009; Ferrari & Comelli, 2016; Krueger & Tian, 2004). Due to these advantages, the mixed effects models have become a golden standard in psycholinguistic research (Baayen et al., 2008; Barr, Levy, Scheepers, & Tily, 2013), and are finding their way in the field of memory research (e.g., Friedman, McGillivray, Murayama, & Castel, 2015; Murayama, Sakaki, Yan, & Smith, 2014), but also in other fields of psychology, such as social psychology (e.g., Judd, Westfall, & Kenny, 2012).

It should be noted that there are other potential methods that could be observed as alternative to ANOVA in the analysis of binary data, such as Signal detection theory (Wixted, 2007) or Diffusion model (Ratcliff & McKoon, 2008; Ratcliff, Smith, Brown, & McKoon, 2016). However, these models fit better with data from recognition memory tasks, whereas in this paper we focus on data from recall tasks, namely the free recall task and the cued recall task.

The goal of this paper was to compare the results of ANOVA with the results of the binary logistic regression and the mixed logit models. For that purpose, the data from a previously reported study were used (Popović Stijačić & Filipović Đurđević, 2015). In that study, the proportions of the correct answers were used as the dependent variable values and two ANOVA analyses were conducted – one by-item and one by-participant.

Having in mind that all three types of the analyses were to be run over the same dataset, and that the observed effects were fairly convincing, we expected to observe a similar pattern of results as originally reported by Popović Stijačić and Filipović Đurđević (2015). However, we expected to observe differences with respect to the bias, the standard error, and the confidence intervals of the estimates from the three methods. In order to capture these differences, we performed bootstrap analysis. The main idea of bootstrapping was to calculate parameter estimates, their bias, their standard errors and confidence intervals (CIs) based on a large number of samples collected from an original sample with replacement (Banjanović & Osborne, 2016; Davison, Hinkley, & Young, 2003; Efron, 2000; Purić & Opačić, 2013). The bias of an estimate reveals an accuracy of an estimate, i.e., it represents the difference between an average of the bootstrap estimates and original estimate (obtained in the original analysis). The standard error (*SE*) and CIs represent the variation of the bootstrap estimates around the estimate. Because the *SE* of the bootstrap estimate is tightly related to CIs, we focused only on CIs. This decision is also rooted in one of the six principles of the bootstrap CIs described by Banjanović and Osborne (2016). According to these authors, if a researcher is focused on the utility of the estimates, the usage of the bootstrapped CIs would be the most informative. For example, if the standard error of the bootstrap estimate was large, the CIs would have been wider, and hence, less reliable. In brief, the most reliable estimate would be the one with the narrowest intervals, and the most accurate would be the one with the bias closest to zero.

In this research, we hypothesized that ANOVA estimates would be least accurate and least reliable compared to the estimates obtained in the binary logistic regression and the mixed logit models. In other words these estimates would be the most biased and with the widest bootstrapped CIs. This expectation was based on the fact that the application of ANOVA over proportions violates two assumptions of the linear models, as previously discussed (Jaeger, 2008). The remaining two analyses (from the generalized linear models family) were specifically designed for modelling binomial data, hence parameter estimates should be less biased and should be less variable with respect to CIs (i.e., should be more accurate and more reliable). Finally, within generalized linear models family, it was hypothesized that, compared to logistic regression, parameter estimates of the mixed logit model should produce even narrower CIs (be more reliable), due to its power to model random effects (Clark & Linzer, 2015; Gelman & Hill, 2007). However, according to Clark and Linzer (2015), introduction of random effects in the model could potentially produce bias in fixed effects estimates. According to this, we could expect more biased estimates in the mixed logit models than in logistic regression models.

Method

Participants

From a total of 91 psychology students, 44 were recruited for a free recall task and 47 were recruited for a cued recall task. All participants were native speakers of Serbian language with a normal or corrected to normal vision.

Stimuli

The study list contained 41 noun pairs; eight noun pairs represented fillers, with regard to the primacy and recency effect (Glanzer & Cunitz, 1966; Murdock, 1962). The rest of 33 noun pairs were divided in three groups. The first group contained concepts that could be experienced with many sensory modalities (three to five; e.g., *orange–peach*), the second group contained concepts that could be experienced with few sensory modalities (one or two; e.g., *needle–sting*), and the third group contained abstract concepts (ones that cannot be perceptually experienced; e.g., *aggression–violence*). All three groups of noun pairs were matched for word familiarity, word length, and logarithm of the lemma frequency (Kostić, 1999). The groups that contained concrete nouns were additionally matched for concreteness and visual perceptual strength. The cue and the target in all noun pairs were associatively or semantically related, according to the ratings of additional 20 participants.

Design

This study was arranged as 2 x 3 mixed factorial design (Figure 1), with the task as between-participant factor and the number of sensory modalities as within-participant factor. Dependent variable in the case of ANOVA analysis was the proportion of correct responses; on the other hand, in the logistic regression and the mixed logit model analysis, we used raw data instead of proportions. In other words, dependent variable was binary coded as 1 *for correctly reproduced word pair* and 0 *for false*.

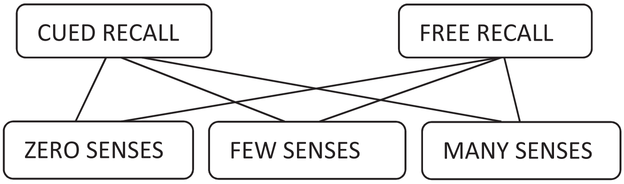

Figure 1. Scheme of research design.

Procedure

The participants were divided into two groups. Each group was presented with the same stimuli list, but approximately half of participants took part in the free recall, and another half in the cued recall task. The stimuli were presented by using a video beam projector. Duration of each stimulus was eight seconds, preceded by a fixation cross for one second. Participants were instructed to read carefully word pairs (the recall test was not mentioned). After the stimuli presentation, the experimenter distributed either a blank sheet of paper for the free recall task, or a sheet of paper containing the list of cues (the first words in word pairs) for the cued recall task. The time for the reproduction was limited to five minutes.

Data analysis

The data were analyzed in R statistical software (R Core Team, 2012). We used the stats package (R Core Team, 2012) when performing ANOVA and the binary logistic regression analysis. The mixed logit models were performed by relying on lme4 package (Bates, Maechler, Bolker, & Walker, 2015). Additionally, some functions from the rms package were applied as well (Harrell, 2015). The full code of our analyses is avaliable at https://github.com/milipopo/mix-and-bootstrap and https://osf.io/7n9ez/.

In the second part of the research, bootstrapping was applied to test for the accuracy (bias) and reliability (CIs) of the obtained estimates: *F* test for ANOVA, *z* test for logistic regression and *t* test for mixed logit model (Davison et al., 2003; Davison & Kuonen, 2002; Efron, 2000). By virtue of this method, a large number of new samples from the original sample were created, all of the same size as the original one. In order to obtain the nonparametric bootstrap estimates in all analyses, the boot package was utilized (Canty & Ripley, 2015). Following recommendations of Davison and MacKinnon (2000), we decided, based on pilot analyses, to take 2000 samples for the calculation of the standard errors and the bias and 10000 samples for the calculation of the confidence intervals.

Results

ANOVA over the proportions of correct answers

Two separate mixed ANOVA analyses were run – one by participants and one by stimuli. The aov function of the R programme was used for ANOVA calculation.

By participant analysis – *F1* test. The mixed ANOVA analysis was applied with the number of sensory modalities as the within factor, and the task as the between factor (R code avaliable at https://github.com/milipopo/mix-and-bootstrap and https://osf.io/7n9ez/). We observed the main effect of the task, $F1(1, 89) = 57.57$, $p < .001$, $\eta_p^2 = .39$, as well as the main effect of the number of sensory modalities, $F1(2, 178) = 24.106$, $p < .001$, $\eta_p^2 = .21$. The task by the number of sensory modalities interaction did not reach statistical significance, $F(2, 178) = .177$, $p = .838$, $\eta_p^2 = .002$. In other words, participants were significantly better in the cued recall task and in memorising concepts that could be perceptually experienced compared to the abstract ones.

By stimuli analysis – *F2* test. In this analysis, the task was included as the within factor and the number of sensory modalities as the between factor. As with by-participant analysis, we observed the main effects of task and the number of sensory modalities, but no interaction: $F2(1, 30) = 181.94$, $p < .001$, $\eta_p^2 = .86$; $F2(2, 30) = 6.33$, $p < .01$, $\eta_p^2 = .30$; $F(2, 30) = .213$, $p = .81$, $\eta_p^2 = .01$. Again, reproduction was better in cued recall and for concepts that could be perceptually experienced.

**Min F` calculation.** Finally, we calculated *min F`* test, according to the Clark`s (1974) recommendations as *min F`(i, j) = F1 * F2* / (*F1* + *F2*). If *F1* has *n* and *n1* degrees of freedom, and *F2* has *n* and *n2* degrees of freedom, *i* = *n* and *j* is represented by the nearest integer of the following expression:

$j = (F1 + F2)^2 / (F1^2/n2 + F2^2/n1)$. When the replacement of all relevant values was done, we obtained significant min *F`* statistic, for both task and the number of modalities: min *F`* task(1, 119) = 43.73, $p < .01$ and *min F`* NoM(2, 47) = 5.01, $p < .05$.

Logistic regression analysis

In this analysis, the long data format was used, and the dependent variable was binary coded. Two models were formulated – the first one tested only main effect of the task and the number of sensory modalities, and the second one included their interaction.

The first and the second model were created to test if the task by number of sensory modalities interaction contributed significantly to prediction of the correct responses. This was achieved by comparing data fits obtained in separate models. Table 1 shows the estimates of the logistic regression coefficients and the fit indices for each model.

Table 1
Comparison of the coefficients and the fit indices of the two binary logistic regression models

Model	*b*	*SE*	*z*	*p*	Fit indices
Model 1					AIC = 3779.1
Intercept (free recall, zero senses)	-1.39	.09	-16.25	***	Pseudo R^2 = .13
Task: cued recall	1.24	.08	15.68	***	
Number of senses: few	0.43	.097	4.46	***	
Number of senses: many	0.67	.096	6.99	***	
Model 2					AIC = 3781.6
Intercept (free recall, zero senses)	-1.46	.12	-12.58	***	Pseudo R^2 = .13
Task: cued recall	1.36	.12	9.31	***	
Number of senses: few	0.49	.16	3.19	**	
Number of senses: many	0.81	.15	5.4	***	
Cued recall*Few senses	-0.1	.20	-0.48	.63	
Cued recall*Many senses	-0.24	.20	-1.22	.22	

Legend. *** $p < .001$, ** $p < .01$, * $p < .05$. Reference category for the task was the free recall, and the reference category for the number of senses was the zero senses.

As Table 1 shows, the results were fully comparable to those obtained with the ANOVA – the effects of the task and the number of sensory modalities was observed, but not their interaction. Comparison of nested models revealed that no significantly better data fit was obtained with the second model (interaction model): $\Delta\chi^2 = 1.54$, $df = 2$, $p = .46$.

Mixed logit model analysis

The advantage of the mixed logit model analysis compared to the binary logistic regression is that it allows for modelling of the random effects. We started by fitting the model with the full random structure, as suggested by Barr et al. (2013). However, the model was not able to converge, as was not any of the models that included random slopes. Therefore, we switched to the strategy of building the model bottom-up, as suggested by Barr et al. and compared models with respect to goodness of fit measure, as suggested by Baayen (2008) and Bates, Kliegl, Vasishth, and Baayen (2015). We did so for the model that were able to converge.

In order to test for the same fixed effects of task and number of sensory modalities, while at the same time including the random effects of participants and items, five models were created, as presented in Table 2. The first model contained the random intercepts of participants, the second one contained the random intercepts of items, and the third model contained the random intercepts of both participants and items. The comparison of the first three models was to inform us whether the simultaneous inclusion of both random effects in the model was justified. The fourth model contained the fixed effects of the task and the number of senses, and the fifth model additionally included their interaction (R code avaliable at https://github.com/milipopo/mix-and-bootstrap and https://osf.io/7n9ez/). The comparison of the fourth and the fifth model was to inform us if the inclusion of the interaction term was justified.

Table 2
Model schema for mixed logit models analysis

Model	Model schema
Model 1	Response ~ 1 + (1|participant)
Model 2	Response ~ 1 + (1|item)
Model 3	Response ~ 1 + (1|item) + (1|participant)
Model 4	Response ~ Task + Number of sensory modalities + (1|item) + (1|participant)
Model 5	Response ~ Task * Number of sensory modalities + (1|item) + (1|participant)

The fit indices of the tree models which were created to test for the random effects, as well as, results of their comparisons are presented in Table 3. These analyses revealed that the third model, the model with both random intercepts of participants and items was the model with the best fit indices (the smallest value of AIC and BIC, and the largest value of Log likelihood), compared to models that contained only one random intercept. This indicated that both by-item and by-participant random variance significantly contributed to prediction of responses.

Table 3
Comparison of the fit indices of the three models that include only random effects

	df	AIC	BIC	logLik	Deviance	$\Delta\chi^2$	*p*
Model 1	1	3672.9	3685	-1834.5	3668.9		
Model 3	2	3569.4	3587.4	-1781.7	3563.4	105.56	***
Model 2	1	4007.8	4019.8	-2001.9	4003.8		
Model 3	2	3569.4	3587.4	-1781.7	3563.4	440.45	***

Legend. df – degrees of freedom; AIC – Akaike information criterion; BIC – Bayesian information criterion; log Lik – Log likelihood; χ^2 – chi square; *** $p < .001$.

For the fourth and the fifth model, the estimates of the random effects are shown in Table 4.1, whereas Table 4.2 contains the estimates of the coefficients for the fixed effects included in these models and the fit indices. As it could be seen from Table 4.2, the same pattern of results was observed as in the previous two analysis: significantly better recall in the cued recall task, significantly better recall of the concrete noun pairs than the abstract ones, and the absence of interaction.

Table 4.1
The variance and the standard deviation of the random effects of the fourth and the fifth model

Model	Source of the variability	Variance	Std. Deviation
Model 4	participant (intercept)	0.68	0.82
	item (intercept)	0.20	0.44
Model 5	participant (intercept)	0.67	0.82
	item (intercept)	0.20	0.44

Table 4.2
The estimates of the coefficients of the fixed effects and the fit indices for the fourth and the fifth model

	Estimate	*SE*	*t*	*p*	Fit indices
Model 4					
Intercept (free recall, zero senses)	–1.59	.20	-7.77	***	
Task: cued recall	1.44	.19	7.44	***	AIC = 3519.8
Number of senses: Few	0.50	.21	2.30	*	BIC = 3555.8
Number of senses: Many	0.79	.22	3.65	***	logLik= -1753.9
Model 5					
Intercept (free recall, zero senses)	-1.64	.22	-7.46	***	
Task: cued recall	1.52	.23	6.54	***	AIC: 3522.7
Number of senses: Few	0.52	.25	2.07	*	BIC: 3570.8
Number of senses: Many	0.90	.25	3.64	***	logLik= -1753.3
Cued recall*Few senses	-0.03	.21	-0.12	.90	
Cued recall*Many senses	-0.20	.21	-0.94	.35	

Legend. *** $p < .001$, ** $p < .01$, * $p < .05$. Reference category for the task was the free recall, and the reference category for the number of senses was the zero senses.

The first part of this research showed that all three analyses gave the same pattern of the results. The aim of the second part of the study was to explore which of the estimates were the most reliable ones, and hence the most appropriate for binomial data. To answer this question, the bootstrap analysis of the given estimates was performed.

Comparison of the estimates of the three methods – the bootstrap analysis

Bootstrapping of the ANOVA estimates. In Table 5, the bias, the standard error and the confidence intervals of the *F* estimates from by-participant analysis are given. In Table 6, the same markers are presented for the by-stimuli analysis. We observed the same tendencies in bootstrap analyses for the two ANOVAs. As presented in the two tables, the *F* estimate for the main effect of the task had the largest bias and the largest standard error, and hence the widest confidence intervals. The *F* estimate for the main effect of the number of sensory modalities had somewhat smaller bias than the *F* estimate for the task; furthermore, it had smaller standard error and narrower CIs. It is interesting to point out that the *F* estimates for the interaction (which was not statistically significant) had the smallest bias and the smallest standard error, and consequently the narrowest CIs in absolute values (however, see discussion on relative bias and relative CI below). Figure 2 depict the distributions and the quantile diagrams of the bootstrap *F* estimates for the three effects, for by-subjects and by-stimuli analyses as well. As illustrated in Figure 2, the distributions of the *F* estimates for the task and the number of sensory modalities were slightly positively skewed, as expected. The distribution of the *F* estimates for the interaction was severely skewed to the right and it had the shape of the distribution for the rare events.

Table 5
The bootstrap analysis of the F estimates in by-subject ANOVA analysis

				95 % Confidence intervals		
Effect	Original estimate	Bias	*SE*	Basic	Percentile	Bias corrected
Task	57.57	2.85	18.83	[10.29, 84.56]	[30.59, 104.86]	[30.10, 102.90]
NoM	24.12	1.75	8.20	[4.29, 36.11]	[12.10, 43.93]	[10.57, 40.98]
Interaction	0.18	0.99	1.19	[-4.09, 0.33]	[0.03, 4.44]	[0.00, 0.64]

Table 6
The bootstrap analysis of the F estimates in by-stimuli ANOVA analysis

				95 % Confidence intervals		
Effect	Original estimate	Bias	*SE*	Basic	Percentile	Bias corrected
NoM	6.33	1.78	4.02	[-5.42, 10.35]	[2.32, 18.09]	[1.54, 13.90]
Task	181.94	31.83	67.15	[-16.70, 253.10]	[110.80, 380.50]	[85.90, 300.70]
Interaction	0.12	1.19	1.57	[-5.40, 0.22]	[0.03, 5.65]	[0.00, 0.39]

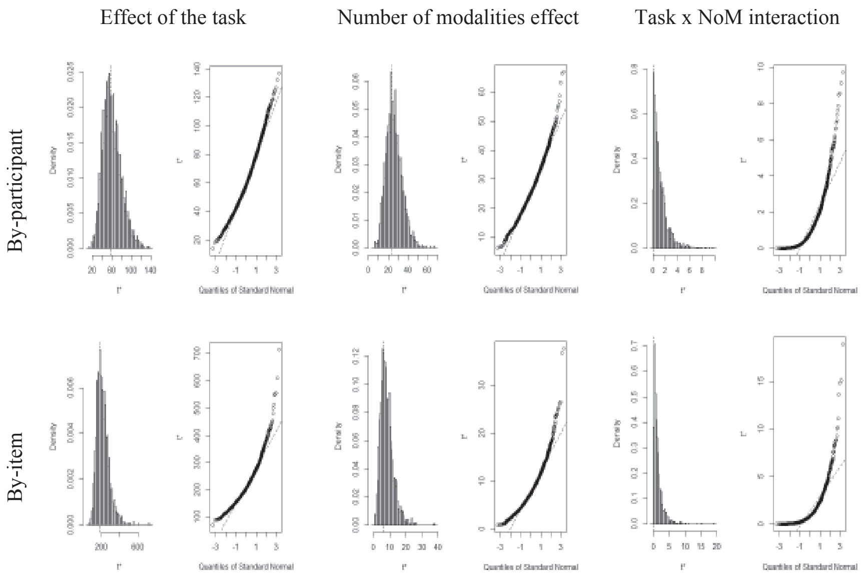

Figure 2. The distribution of the *F* estimates for the main effect of the task, the number of sensory modalities and their interaction (in by-participant and by-item analyses).

Bootstrapping of binary logistic regression. The code for the bootstrap analysis for the logistic regression was adopted from Hossain and Khan (2004, see code at https://github.com/milipopo/mix-and-bootstrap and https://osf.io/7n9ez/). In Table 7, the original estimate, the bias, *SE* and CIs are presented for the main effects of the task and the number of sensory modalities. As with the previous analysis, 2000 samples were extracted for calculation of *SE* and the 10000 samples for calculation of CIs. The bias for all estimates was approximately zero, and its value was the smallest for the effect of the task. The standard errors for all *z* estimates were around one. The confidence intervals were similar for all *z* estimates and compared to the ANOVA bootstrap estimates, they were narrower, with range of approximately 3.5. In Figure 3, the distribution and the quantile diagrams are shown. It is notable that the distributions of *z* statistics were symmetrical and approximately normal. It could be concluded that compared to the ANOVA-based *F* estimates, the *z* estimates were less biased and less variable, and consequently more stable and more reliable.

Table 7
The bootstrap analysis of z statistic from the logistic regression

				95 % Confidence intervals		
Effect	Original estimate	Bias	*SE*	Basic	Percentile	Bias corrected
Cued: Free	15.68	0.01	0.89	[13.98, 17.47]	[13.89, 17.37]	[13.91, 17.39]
Few: Zero	4.46	-0.05	1.01	[2.59, 6.44]	[2.47, 6.32]	[2.54, 6.40]
Many: Zero	6.99	-0.03	0.98	[5.13, 8.91]	[5.07, 8.85]	[5.10, 8.88]

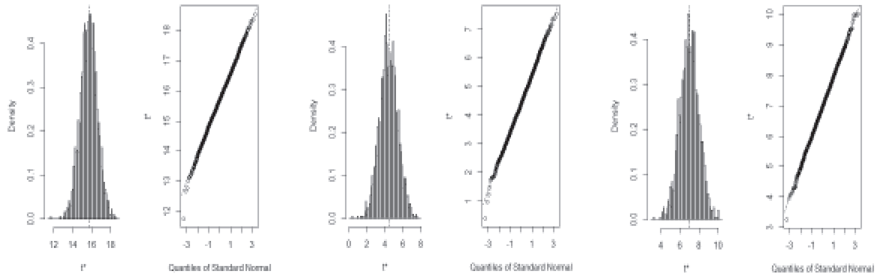

Figure 3. The distribution of the *z* estimates.

Legend. The left panel represents the distribution of the z estimates for the task effect, the middle panel shows the distribution of the *z* estimates for the main effect of the number of modalities for the few:zero effect and the right panel represents the distribution of the *z* estimates for the main effect of the number of modalities for the many: zero effect.

Bootstrapping of the mixed logit model. In Table 8, the results of the bootstrap analysis of the fixed effects from the fifth model are presented, as previously was found that this model has best fit indices (Table 3). Compared to bootstrap analysis of the logistic regression estimates, the bias of the *t* statistic was bigger than the bias of the *z* statistic. On the other hand, the *SE* of the *t* statistic was smaller than *SE* of the *z* statistic, and hence the CIs were narrower for the estimates from the mixed logit model analysis. In Figure 4, the distributions of the *t* estimates were given, as well as the quantile diagrams. It is observable that the distributions were symmetric and approximately normal.

Table 8
The bootstrap analysis of t statistic from the mixed logit model

				95 % Confidence intervals		
Effect	Original estimate	Bias	*SE*	Basic	Percentile	Bias corrected
Cued: Free	7.44	-0.61	0.46	[7.14, 8.92]	[5.97, 7.75]	[7.15, 8.99]
Few: Zero	2.30	-0.21	0.48	[1.55, 3.39]	[1.20, 3.04]	[1.63, 3.65]
Many: Zero	3.65	-0.32	0.51	[2.92, 4.92]	[2.39, 4.39]	[3.02, 5.35]

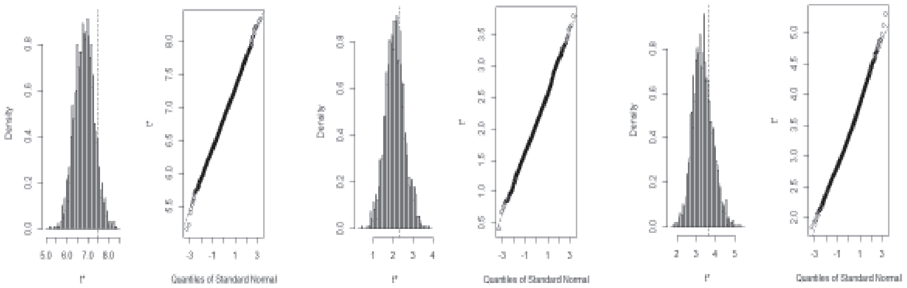

Figure 4. The distribution of the *t* estimates.

Legend. The left panel represents the distribution of the t estimates for the task effect, the middle panel shows the distribution of the *t* estimates for the main effect of the number of modalities for the few:zero effect and the right panel represents the distribution of the *t* estimates for the main effect of the number of modalities for the many: zero effect.

Calculation of relative bias and relative CI. Because the absolute values of bias can depend on the order of magnitude of the estimated statistic, in order to make the three kinds of estimates fully comparable, we also calculated the relative bias, according to expression *relative bias = (bias / original estimate) * 100.* Relative bias depicts the accuracy of estimation in terms of percent relative to the original estimate. In Table 9.1, the comparison of relative bias estimates are given for all conducted analyses. It can be noticed that relative bias estimates are the lowest for logistic regression coefficients, followed by those for *F1* statistics (*F* estimates from by-participant analysis), than the estimates for mixed logit model coefficients, and the highest values are recorded for *F2* bootstrap estimates (*F* estimates from by-item analysis). These results show that the greatest variance and noise come from diversity of items. This analysis clearly marked by-item ANOVA as the least accurate and logistic regression as the most accurate analysis. Although by-participant ANOVA and mixed-effect logit model were of comparable accuracy, it should be kept in mind that, unlike by-participant ANOVA, mixed effect logit model also included items which were attested as the higher source of imprecision in coefficient estimation.

In a similar manner, we introduced relative CI, according to expression:

$$\text{Relative CI} = \{[(\text{CI}_{\text{upper limit}} - \text{CI}_{\text{lower limit}}) / 2] / \text{original estimate}\}*100$$

Upper and lower limits of CI were taken from the CI bias corrected estimates. Relative CI reveals to what extent the original estimate could change in 95% of experiments, either in positive, or negative direction. For example, relative CI of 50% would indicate that the original estimate could be up to 50% larger, or 50% smaller relative to its originally estimated value (i.e., it could change up to ±50% of its value). As illustrated in Table 9.2, the 95% CI bounds of the expected fluctuations around ANOVA based estimates are approximately 60% of the originally estimated values of coefficients (or larger – e.g., the coefficient for the NoM effect could be almost twice as large, or very close to zero). At the same time, the coefficients estimated in logistic regression and in mixed logit models are expected to fluctuate no more than 11–43% of their original value. This part of analysis clearly distinguished both variants of ANOVA from both remaining approaches, with latter being more reliable in coefficient estimation.

Table 9.1
Summary of the relative bootstrap estimates: relative bias

Effect	ANOVA by subjects	ANOVA by items	Logistic regression	Mixed logit model
Many: Zero	/	/	0.43	8.8
Few: Zero	/	/	1.18	9.13
NoM	7.26	28.12	/	/
Task	4.95	17.49	0.06	8.2

Table 9.2
Summary of the relative bootstrap estimates: relative CI

Effect	ANOVA by subjects	ANOVA by items	Logistic regression	Mixed logit model
Many: Zero	/	/	27.07	31.92
Few: Zero	/	/	43.27	43.91
NoM	63.04	97.63	/	/
Task	63.23	59.03	11.1	12.37

Discussion

The main goal of this research was to explore which of the three statistical models was the most appropriate and the most efficient statistical model for analysing the binary coded responses in the memory tasks: ANOVA over proportion of accurate responses, logistic regression, or mixed logit model analysis over the binary coded data. Similar work was presented for categorical responses collected in language comprehension and language production tasks (Jaeger, 2008). However, to our best knowledge, this is the first research addressing this issue for data collected in memory tasks such as cued recall and free recall in paired associate learning. Therefore, in the first part of the research, we applied the strategy previously presented by Jaeger (2008), that is, we compared the results of the three different statistical methods. In the second part we investigated the problem in more depth by utilizing the bootstrap analysis. The motivation for this study had both methodological and conceptual aspects. The methodological aspect concerned questioning the appropriateness of GLM methods (such as ANOVA) when analyzing proportions as dependent variables (Baayen, 2012; Ferrari & Comelli, 2016; Jaeger, 2008), whereas the conceptual aspect referred to the issue of the language as the random effect (Baayen, 2012; Clark, 1974; Coleman, 1964; Quené & van der Bergh, 2008; Raaijmakers et al., 2008).

In the first part of the study, we presented in brief the conceptual and methodological core of the previously reported study (Popović Stijačić & Filipović Đurđević, 2015) which provided data for the current study. In the initial study, following the tradition in this research area (Marschark & Hunt, 1989; Marschark & Surian, 1992; McDougall & Pfeifer, 2012; Paivio, Clark, & Khan, 1988; Paivio, Walsh, & Bons, 1994; Tse & Altarriba, 2009; Schwanenflugel, Akin, & Luh, 1992; Schwanenflugel, Harnishfeger, & Stowe, 1988) the ANOVA analysis was run over the proportions of the correct answers. Therefore, we started by rerunning the ANOVA analysis, followed by two statistically more appropriate analyses for the binary coded data. The first one was the logistic regression (binary logit model), and the second one was the mixed logit model analysis, both belonging to family of generalized linear models (Agresti, 2002; Baayen, 2012; Jaeger, 2008). The binary logistic regression is the statistical method designed for binomial data. However, it does not allow for modelling of random effects. If a researcher needs to include random effect in the model, as is the case in psycholinguistics and language-based memory tasks, mixed logit models have the advantage over the logistic regression. Capturing all the variation that originates in participants or stimuli allows for greater precision in data modelling (Clark & Linzer, 2015). As the ANOVA analysis of our data revealed solid main effects of the task and the number of sensory modalities, we hypothesized that the similar results would be observed with the application of more appropriate analyses. As predicted, the same pattern of effects was observed with both of generalized linear models analyses. Reproduction was more accurate in the cued recall task, compared to the free recall. Furthermore, reproduction was better for the concepts that could be perceptually experienced (concrete words). None of these analyses revealed the effect of interaction. The observed results were in accordance with the hypothesis that (in this case) all three statistical methods would give similar pattern of results. However, based only on this part of research, we could not discern which statistical method was the most appropriate for analysis of the binomial data. Although it would be possible to estimate effect size for each of the three analyses, it would not be possible to compare them in a straightforward manner, as they approach the problem of multiple items per multiple participants differently (as discussed in Brysbaert & Stevens, 2018). Therefore, in order to compare the three methods, we conducted the bootstrap analysis.

In the second part of the study, using nonparametric bootstrap (Davison et al., 2003; Efron, 2000), we calculated the bootstrap estimates, their bias, standard errors, and confidence intervals. Because ANOVA is considered inappropriate for the binomial data, we hypothesized that their estimates (bootstrapped *F* statistic) would have the highest bias, highest *SE*s, and consequently the widest CIs. For the two methods from generalized linear models, we predicted that the estimates of the mixed logit model would be more reliable compared to the estimates of the binary logistic regression, as it allows for modelling of the random effects (with the risk of being more biased, as a consequence). As predicted, ANOVA bootstrap estimates had the highest bias, the highest *SE*s, and the widest CIs compared to the two generalized linear models methods. In other words, ANOVA based estimates were the least accurate and the least reliable, which is of importance for the replication potential of the observed effect. In spite of the fact that in this particular case (with strong effects) ANOVA results were confirmed, the chances of obtaining the accurate estimates were not very large. Additionally, the bias of ANOVA based measures was always above zero, indicating that ANOVA was overestimating the coefficient values (whereas the remaining two methods tended to underestimate). In other words, if we use ANOVA to analyze proportions, the probability of Type I error grows, and consequently our conclusions based on it could be very misleading (Ferrari & Comelli, 2016; Jaeger, 2008).

Bootstrap analysis showed that both methods from generalized linear models were equally good, in terms of accuracy and reliability. However, although both methods had the bias around zero, the bootstrap estimates of the binary logistic regression were slightly less biased than the estimates of the mixed logit model. Such pattern of results was expected, because it was found that introduction of random effects in the model could consequently give more biased fixed effect estimates (Clark & Linzer, 2015). This difference became more pronounced when we expressed the bias relative to the magnitude of the original estimate. Although the estimates of the mixed logit model had lower standard errors, and consequently narrower confidence intervals, expressing them in relative terms revealed that the two were of comparable variability. This implies that the estimates of the logistic regression were more accurate and equally reliable compared to estimates of the mixed logit model. We believe that the main reason for this is the lack of random slope specifications in the mixed model (at least for some of the effects), which was due to lack of convergence of these models. However, one should keep in mind that the observed differences between logistic and mixed effects regression were very small. Distributions of estimates of both methods were symmetrical and approximately normal, which suggested that these estimates were not biased. Consequently, these two methods should be applied to binomial data, without the fear of greater probability of making Type I or Type II errors.

Conclusion

This research answered both the methodological and the conceptual aspects of the initial research question. Methodologically, we demonstrated that ANOVA should be avoided when analyzing binomial data, such as the data from memory tasks. Conceptually, we contributed to the "Language-as-fixed-effect-fallacy" debate (Clark, 1974). The language can be treated as the random factor, and we have an easy tool for such analysis, that tool being the mixed logit model analysis. However, if comparison of nested models reveals that there is no necessity for inclusion of random factor, or the data do not support the convergence of the model that includes random slopes, a researcher can opt for a simpler analysis, namely binary logistic regression.

References

Agresti, A. (2002). *Categorical data analysis* (Second ed.). New York, NY: John Wiley & Sons.

Baayen, R. H. (2012). Mixed-effects models. In A. C. Cohn, C. Fougeron, C., & M. K. Huffman (Eds.), *Handbook of Laboratory Phonology* (pp. 668–677). Oxford: Oxford University Press.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge, UK: 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.

Banjanović, E. S., & Osborne, J. W. (2016). Confidence Intervals for Effect Sizes: Applying Bootstrap Resampling. *Practical Assessment, Research & Evaluation, 21*(5), 1–20.

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

Bates, D., Kliegl, R., Vasishth, S., & Baayen, R. H. (2015). Parsimonious mixed models. Available from arXiv: 1506.04967 (stat.ME).

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). *lme4: Linear mixed-effects models using Eigen and S4.* R package (Version 1.1–9) [Computer software]. Retrieved from https://CRAN.R-project.org/package=lme4.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J-S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution, 24*(3), 127–135. https://doi.org/10.1016/j.tree.2008.10.008

Brewer, G. A., & Unsworth, N. (2012). Individual differences in the effects of retrieval from long-term memory. *Journal of Memory and Language, 66,* 407–415.

Brysbaert, M., & Stevens, M. (2018). Power Analysis and Effect Size in Mixed Effects Models: A Tutorial. *Journal of Cognition, 1*(1). http://doi.org/10.5334/joc.10

Clark, H. H. (1973). The Language-as-Fixed-Effect Fallacy: A Critique of Language Statistics in Psychological Research. *Journal of Verbal Learning and Verbal Behavior, 12,* 335–359.

Clark, T. S., & Linzer, D. A. (2015). Should I Use Fixed or Random Effects? *Political Science Research and Methods, 3*(2), 399–408.

Coleman, E. B. (1964). Generalizing to a language population. *Psychological Reports, 14,* 219–226.

Davison, A. C., Hinkley, D. V., & Young, G. A. (2003). Recent Developments in Bootstrap Methodology. *Statistical Science, 18*, 141–157.

Davison, A. C., & Kuonen, D. (2002). An introduction to the bootstrap with applications in R. *Statistical Computing and Statistical Graphics Newsletter*, *13*(1), 6–11.

Davidson, R., & MacKinnon, J. G. (2000). Bootstrap Tests: How Many Bootstraps? *Econometric Reviews 19*(1)*,* 55–68.

Efron, B. (2000). The Bootstrap and Modern Statistics. *Journal of the American Statistical Association, 95,* 1293–1296.

Ferrari, A., & Comelli, M. (2016). A comparison of methods for the analysis of binomial clustered outcomes in behavioral research. *Journal of Neuroscience Methods, 274,* 131–140.

Friedman, M. C., McGillivray, S., Murayama, K., & Castel, A. D. (2015). Memory for Medication Side Effects in Younger and Older Adults: The Role of Subjective and Objective Importance. *Memory & Cognition, 43*(2), 206–215. http://doi.org/10.3758/s13421-014–0476–0

Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multi-level/Hierarchical Models.* Cambridge: Cambridge University Press.

Glanzer, M., & Cunitz, A. R. (1966).Two Storage Mechanisms in Free Recall. *Journal of Verbal Learning and Verbal Behaviour, 5*(4), 351–360.

Harrell, F. E. (2015). Package rms: Regression Modelling Strategies. R package (Version 4.5–0) [Computer software]. Retrieved from https://cran.r-project.org/web/packages/rms/rms.pdf

IncTse, C-S., & Altarriba, J. (2009). The word concreteness effect occurs for positive, but not negative, emotion words in immediate serial recall. *British Journal of Psychology, 100,* 91–109.

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.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103*(1), 54–69. https://doi.org/10.1037/a0028347

Kostić, Đ. (1999). *Frekvencijski rečnik savremenog srpskog jezika [Frequency dictionary of contemporary Serbian language]*. Beograd: Institut za eksperimentalnu fonetiku i patologiju govora i Laboratorija za eksperimentalnu psihologiju.

Krueger, C., & Tian, L. (2004). A comparison of the general linear mixed model and repeated measures ANOVA using a dataset with multiple missing data points. *Biological Research for Nursing, 6*(2), 151–157. http://dx.doi.org/10.1177/1099800404267682

Marschark, M., & Hunt, R. R. (1989). A reexamination of the role of imagery in learning and memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 710–720.

Marschark, M., & Surian, L. (1992). Concreteness effects in free recall: The roles of imaginal and relational processing. *Memory & Cognition, 20*, 612–620.

Murayama, K., Sakaki, M., Yan, V. X., & Smith, G. M. (2014). Type I Error Inflation in the Traditional By-Participant Analysis to Metamemory Accuracy: A Generalized Mixed-Effects Model Perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*, 1287–1306.

McDougall, S., & Pfeifer, G. (2012). Personality differences in mental imagery and the effects on verbal memory. *British Journal of Psychology, 103,* 556–573.

Murdock, B. B. Jr. (1962). The retention of individual items. *Journal of Experimental Psychology, 62*, 618–625.

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

Paivio, A., Clark, J. M., & Khan, M. (1988). Effects of concreteness and semantic relatedness on composite imagery ratings and cued recall. *Memory & Cognition, 16,* 422–430.

Paivio, A., Walsh, M., & Bons, T. (1994). Concreteness effect on memory: when and why? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1196–1204.

Popović Stijačić, M., & Filipović Đurđević, D. (2015). Uspešnost reprodukcije u zavisnosti od broja čula kojima je moguće iskusiti pojam [Number of sensory modalities through which a concept can be experienced: effect on recall]. *Primenjena psihologija, 8*(3), 335–352. https://doi.org/10.19090/pp.2015.3.335–352.

Purić, D. & Opačić, G. (2013). Poduzorkovanje, samouzorkovanje, postupak „univerzalnog noža" i njihova upotreba u postupcima za statističku analizu multivarijacionih podataka [Resampling, bootstrapping, jackknifing and their use in mulitivariate (statistical) data analyses]. *Primenjena psihologija, 6,* 249–266. http://dx.doi.org/10.19090/pp.2013.3.249–266

Raaijmakers, J., Schrijnemakers, J., & Gremmen, F. (2008). How to Deal with "The Language-as-Fixed-Effect Fallacy": Common Misconceptions and alternative Solutions. *Journal of Memory and Language 41*, 416–426.

R Core Team (2012). R: *A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Ratcliff, R., & McKoon, G. (2008). The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural computation, 20*(4), 873–922.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in Cognitive Science, 20*(4), 260–281.

Schwanenflugel, P. J., Harnishfeger, K. K., & Stowe, R. W. (1988). Context availability and lexical decisions for abstract and concrete words. *Journal of Memory and Language, 27,* 499–520.

Schwanenflugel, P., Akin, C., & Luh, W.-M. (1992). Context availability and the recall of abstract and concrete words. *Memory & Cognition, 20*, 96–104.

Tabachnick, B., & Fidell, L. (2007). *Using Multivariate Statistics.* USA: Pearson Education.

Analiza podataka iz memorijskih zadataka – poređenje ANOVA-e, logističke regresije i mešovitog logit modela

Milica Popović Stijačić[1], Ljiljana Mihić[2] i Dušica Filipović Đurđević[1,2]

[1]Laboratorija za eksperimentalnu psihologiju, Univerzitet u Novom Sadu, Srbija
[2]Odsek za psihologiju, Filozofski fakultet, Univerzitet u Novom Sadu, Srbija

U ovom radu poredili smo binarne ishode tri statističke analize. Kako primena ANOVA-e na proporcijama narušava bar dve klasične pretpostavke linearnih modela, opisane su dve alternative: binarna logistička regresija i mešoviti logit model. Najpre smo poredili efekte dobijene ovim trima metodama na istom skupu podataka, dobijenom u ranijem istraživanju iz oblasti memorije. Rezultati dobijeni ovim trima metodama su bili slični: potvrđeno je postojanje efekata zadataka i broja senzornih modaliteta, ali ne i njihova interakcija. Nakon toga, istražena je efikasnost svakog od metoda korišćenjem procene parametara samouzorkovanjem. Kao što je i predviđeno, procena parametara ANOVA-e samouzorkovanjem je imala veliku pristrasnost i velike standardne greške i, samim tim, široke intervale poverenja. S druge strane, procene parametara binarne logističke regresije i mešovitih logit modela samouzorkovanjem su bile slične – obe metode su imale nisku pristrasnost i niske standardne greške, kao i uske intervale poverenja.

Ključne reči: zadatak vođenog/slobodnog prisećanja, ANOVA, logistička regresija, mešoviti logit modeli, samouzorkovanje

RECEIVED 15.06.2017.
REVISION RECEIVED 06.04.2018.
ACCEPTED 10.04.2018.