Replication of Experiment 2 in "Modeling Distributions of Immediate Memory Effects: No Strategies Needed?" by Beaman, C.P., Neath, I., & Suprenant, A.M. (2008, *Journal of Experimental Psychology: LMC, 34 (1)*, 219 - 239.

Kleinberg, B. & Kunkels, Y.K.
University of Amsterdam
bennettkleinberg@gmail.com, ykkunkels@gmail.com

Introduction

The current study aimed to replicate the findings of Beaman, Neath, and Suprenant (2008). To investigate whether our replication is successful, we will interpret the found test statistics, confidence intervals, and effect sizes in context.The main idea of the original study was to examine whether participants remembered the order of words better when the presented word lists contained only short words compared to lists that only contained long words. The target findings for replication herein were the word-length manipulation effect, and the Word Length Effect (WLE). The former focused on the comparison between the proportion of correct short words versus the proportion of correct long words. The latter is a derivative thereof and focused on the difference between proportions correct for short- and long words, while dividing this difference over the proportion correct for short words. While the original study did find support for the idea that the order of short words is remembered better than with long words, the there was no explicit model which could be compared to an alternative model.

Methods

Power Analysis

Original effect: The first effect was the word-length manipulation which is the proportion correct on short (.48) versus long words (.40), $F(1, 99) = 103.7$, $p < .05$. The second reported effect is the Word Length Effect (WLE) which is

$WLE = (Pcorrect_{short} - Pcorrect_{long}) / Pcorrect_{short}$

and was reported as mean WLE = .16 (SD = .162).

Effect size: We transformed the F statistic of the word-length manipulation into a t-statistic as follows: t = sqrt(F), which is t = sqrt(103.7) = 10.18. This t-statistic was converted into a Cohen's d effect size for paired-samples t-tests using the procedure described by Rosenthal (1991, cited in Lakens, 2013), which is $d = t$ / sqrt(n). The resulting Cohen's d for this effect is then d = 10.18 / 10 = 1.02, 95% CI [0.22, 1.82][1].

Power analysis: The power achieved in the original experiment was computed a posteriori using g*power (Faul, Erdfelder, Lang, Buchner, 2007) and the pwr-package for R (Champely, 2012). With the

[1] The confidence intervals were calculated with the R-package, compute.es (Del Re, A.C., 2014) after computing the effect size for our within-subjects design.

values above the achieved power was indicated as beta = 1 by approximation. Using $d = 1.02$ gave the following sample sizes per condition: N = 10 (.80 power), N = 13 (.90 power), N = 15 (.95 power).[2]

Planned Sample

We assessed our planned sample size and power while acknowledging current OSF guidelines on this point. Therefore, the planned sample consisted of 15 undergraduates. Participants originated from the University of Amsterdam and participated for research credit. Each participant completed two versions of the test - English original and Dutch translation. The order of the versions was randomized. Data collection termination rule was simply reaching 15 participants. The aim was to collect data of exactly 15 participants.

Materials

"The stimuli were 80 long (three to five syllables) and 80 short (one syllable) words from the study by LaPointe and Engle (1990). Sampling from each set was carried out without replacement. "

The experimental task was programmed in Presentation (Neurobehavioral Systems) and the stimuli used were a) the original words from the original article (which in turn were adopted from LaPointe and Engle, 1990) and b) nearly-identical words which were Dutch translations (where possible) from the original ones (the English words were translated into Dutch by a Dutch native speaker). The latter was necessary because our sample consisted mainly of Dutch-speaking undergraduates. The mean word length for the short words was 4.24 letters in both the English original and Dutch translation, and for long words 8.59 and 8.61 letters respectively. In contrast to the original long word list, our translations contained words with two to four syllables. Nevertheless, we expected our planned sample to have good command of the English language. We implemented the different language sets as within-subject variable which was counterbalanced across participants. Each version of the test took approximately 15 minutes, resulting in a total duration of half an hour.[3]

Procedure

"The instructions and procedure were identical to those for Experiment 1 except for the following. Each word was shown in uppercase for 1 s in the middle of the window on a computer screen in 24-point Helvetica font. The list length was eight, and thus eight response buttons were used. Participants received 40 lists, half with short and half with long items. The order of short and long trials was randomly determined for each participant. "

From experiment 1:

[2] One reviewer pointed out the problems with this and referred to Perugini, Gallucci and Constantini, 2014 who discuss the Safeguard Power method. We did not adopt this, however, in order to not deviate from the original OSF replication guidelines.

[3] One OSF reviewer noted that there could be order effects when presenting the different language versions in a counterbalanced way. This issue was then discussed with the first author of the original study who answered that the procedure is fine and we need not adjust the procedure so that we avoid different 'language' orders.

"Participants were informed that we were interested in how accurately they could remember the order in which a series of words had been presented. Each word was shown in uppercase, center justified, for 1.5 s in the middle of the screen in 20-point Helvetica font. After the final word was shown, six response buttons became active and were labeled with the six words in alphabetical order. The participants were asked to indicate the presentation order by clicking on appropriately labeled buttons using the mouse. For example, if they thought the first word was break, they should click on the button labeled break first. If they thought the third word was vote, they should click on the button labeled vote third. Participants received 20 lists, half with dissimilar and half with similar items, and were informed they could take rest breaks at any point. The order of dissimilar and similar trials was randomly determined for each participant. Participants were tested individually, and an experimenter remained in the room to ensure compliance with the instructions."

Analysis Plan

The main analysis consisted of a) comparing the proportion correct for the short words versus the long words, and b) transforming these into a WLE. Following the to-be-replicated study, for a) we used a paired-samples t-test, as in the original study; for b) the WLE was computed as explained in *Power Analysis*. In addition to the original study, however, we compared both the proportion correct and the WLE across language versions. The WLE was compared across language versions using a paired-samples t-test with language as independent variable.
[The design was 2 (language: English vs Dutch) by 2 (word length: short vs long) by 2 (accuracy: correct vs false), so we used logistic regression analysis with accuracy as dependent variable. ***This part of the analysis was not deemed necessary any longer by the authors****]*

Differences from Original Study

The main difference was that our planned sample consisted of Dutch-speaking undergraduates, this is why we translated the stimuli into Dutch as well as using the original English stimuli. The experimental task was programmed in Presentation explicitly for this replication, but we do not expect any differences as we adopted the detailed instructions concerning font, letter size, and stimulus intervals from the original experiment.

(Post Data Collection) Methods Addendum

Actual Sample

For the current replication, our final sample consisted of 15 undergraduate students from the University of Amsterdam who participated for research credit. The mean age was 19.67 years (*SD* = 1.19). Thirteen participants were female and all but one were Dutch native speakers. The mean self-rating in English language proficiency on a scale from 1 (poor English language skills) to 10 (very good English language skills), was 7.5 (*SD* = 1.41). There were no rules for excluding participants from this study and consequently no participant was excluded from the data.

Differences from pre-data collection methods plan

none

Results

Data preparation

All data was collected in a comma separated text file, after which demographical data was separated from task responses. Both data sets were saved and then analysed using R Studio statistical software v0.98.501. The exact R code used for the analysis is available in this project's node on the OSF website.

Confirmatory analysis

The mean proportion of correctly remembered English words in presented order was .58 (*SD* = .11) for short words and .57 (*SD* = .11) for long words, $t(14) = 0.496$, $p = .314$, one-sided, ns. The mean WLE was 0.01 (*SD* = .14, lower quartile = -.445, median = .018, upper quartile = .081, range = -.329 to .219). The distribution of the WLE did not differ significantly from the normal, as indicated by the Shapiro-Wilk test, $W = .946$, $p = .469$, ns.

Exploratory analyses

Besides our confirmatory analysis, we also conducted a number of exploratory analyses aimed to extract additional information from the data.There is some probability that, given the current sample size, the results from the exploratory analysis might be underpowered. Participants did not recall significantly more short words ($M = .62$, $SD = .13$) than long words ($M = .59$, $SD = .13$), $t(14) = 0.986$, $p = .341$, ns. The mean WLE for Dutch words was .028 (*SD* = .262, lower quartile = -.053, median = .074, upper quartile = .177, range = -.698 to .348), with a distribution not significantly different from the normal distribution, $W = 0.8839$, $p = .054$, *ns*.

Furthermore, the overall proportion correct irrespective of word length was .57 (*SD* = .11) for English words and .61 (*SD* = .12), $t(27.81) = -0.787$, $p = .438$, *ns*. Comparing short words ($M = .60$, $SD = .10$) and long words ($M = .58$, $SD = .10$) irrespective of language resulted in a non-significant difference

as well, $t(28) = 0.607$, $p = .549$, *ns*. Using the procedure from Lakens (2013), as described under Power Analysis, the effect size for the word length manipulation was $d = .13$, 95% CI [-0.62, 0.88] for English words and $d = .26$, 95% CI [-0.49, 1.01] for Dutch words. The achieved power was .11 and .16 respectively.

As an overview of how our two-languages design affected the other variables and how the performance on one language related to the other, we examined the intercorrelations of our key variables. Table 2 shows these Pearson correlations.

Table 1. Intercorrelations of the Different Word Versions..

	English short	English long	Dutch short	Dutch long	WLE English	WLE Dutch
English long	.78*					
Dutch short	.35	.14				
Dutch long	.40	.41	.46			
WLE English	.31	-.35	.23	-.09		
WLE Dutch	-.01	-.19	.53*	-.48	.26	
Language proficiency	-.12	.01	.16	.33	-.25	-.23

Note. * = $p < .05$

Noteworthy of the correlations is that only both English short and English long words (r = .78), and Dutch short and the Dutch WLE (r = .53) correlated significantly. The lack of significant correlations between Dutch short and long words, between English and Dutch words, and between English language proficiency and English words performance is an issue of concern and will be discussed below.

Additionally to the proportion correct analysis, we investigated exploratorily how reaction times differed from word to word and between language versions. This allows for another check whether the time to choose a word differed between the two language versions. Table 2 shows the mean reaction times for each of the eight words for both language versions. The reaction time for all words was not significantly different for the two language versions which suggests that there was no bias in language that made the subjects consider the English word order longer than their native Dutch words. Interestingly, each the reaction time for each word differed significantly from the reaction time on the previous word. These results could be expected to be found in the first words but to a lesser extent for the last words where there are less options left. Figure 1 represents these findings graphically and shows the overall trend of decreasing differences between reaction times for the last words. The graph also illustrates that the two language versions did not differ much for the full range of words.

Table 2. Mean Reaction Times in Milliseconds per Chosen Word across Language Versions (SDs in parentheses).

Word	English words	Dutch words
1	4512.733 (1705.539) [a A]	4511.933 (1930.393) [a A]
2	6393.867 (2118.206) [a B]	6432.667 (2157.113) [a B]
3	8769.467 (2817.558) [a C]	8845.067 (2760.628) [a C]
4	10761.07 (3387.966) [a D]	10755.47 (3035.733) [a D]
5	12599.47 (3761.343) [a E]	12998.47 (3382.007) [a E]
6	13915.53 (3949.220) [a F]	14374.93 (3629.971) [a F]
7	14854.87 (4240.857) [a G]	15363.47 (3805.337) [a G]
8	15541.13 (4321.089) [a H]	16124.13 (3833.443) [a H]

Note. Same lowercase letters indicate no significant difference between language versions as calculated with paired-samples t-tests; different uppercase letters indicate a significant difference between two subsequent words with a p-value < .001 as calculated with paired-samples one-sided t-tests.

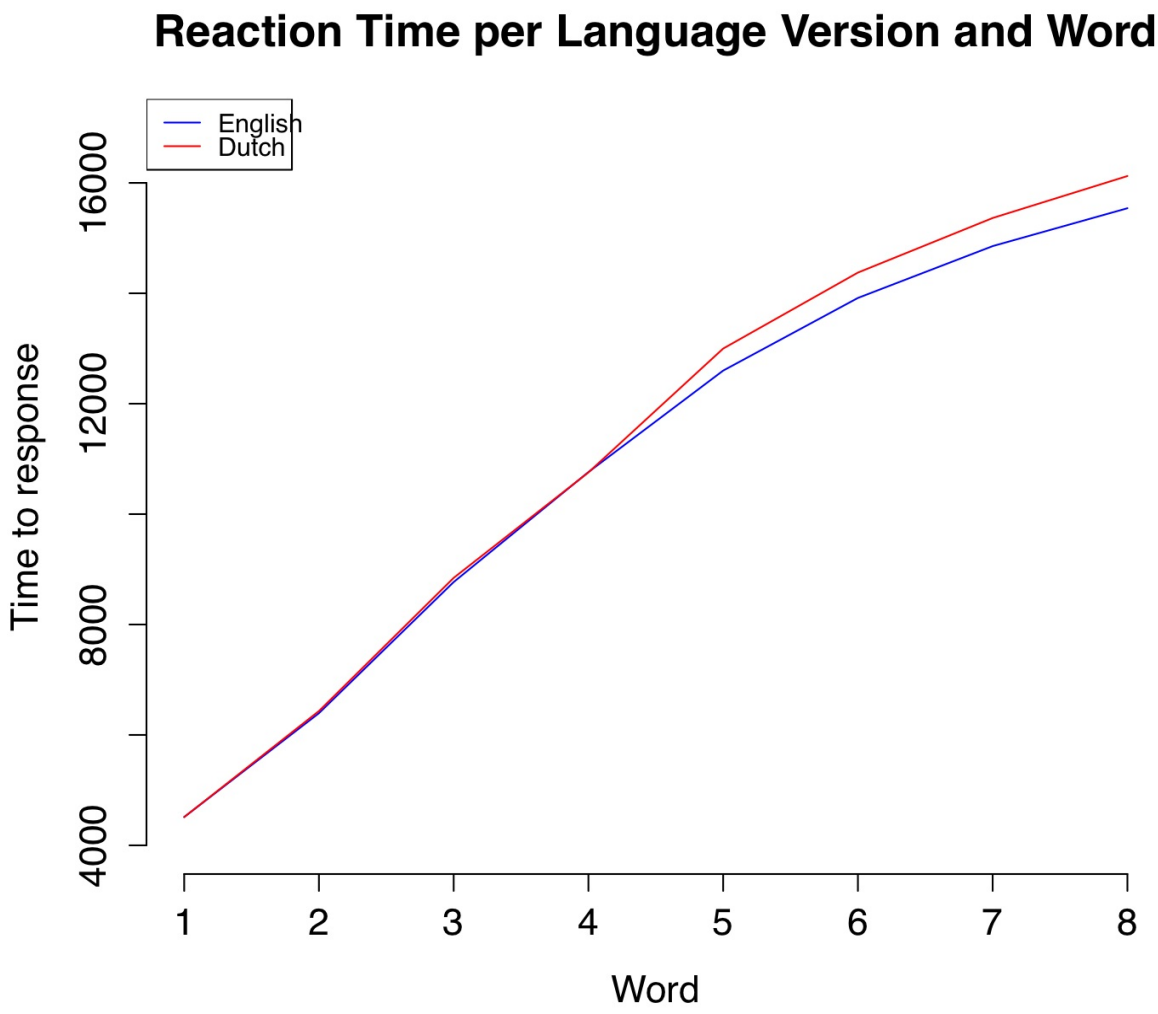

Figure 1. Mean reaction times of word selection per language version irrespective of length.

Discussion

Summary of Replication Attempt

Considering the results from our replication study, we were not able to directly replicate the original findings reported by Beaman et al. (2008). We were unable to replicate the perceived word length effect for proportion correct on English words from the original study. The effect size from the original study was found to be outside of the 95% confidence interval of the effect size of the current replication. Such a result was also found when the effect size from the current study was compared to the 95% confidence interval of the original study. These findings support the notion that our study was unable to

replicate the findings from the original study. But it is not to say that Beaman et. al.'s (2008) results really deviate from what is to be expected in such a setting. The direction of our results, although not significant, did point towards the same direction as results from the original study, as proportion correct was found to be higher for short words. However, it was also found that the effect size for Dutch words did fall within the confidence interval of the original study. There are, of course, limitations to the current study. Although the current replication was designed fully according the OSF guidelines for replication, some scholars might question the sample size, stability of effect size and power[4]. Of course, such concerns are at least partially justified, as effect sizes can be less dependable at small sample sizes, and the notion that the results from the original study are might be an overestimation. Especially our exploratory analyses have to be treated with caution due to the underpowered nature of this replication.

Commentary

The current replication study and exploratory analysis yielded interesting results. For example, in all classes of the current replication, our participants had shown a higher proportion of correct than in the original study. This finding is interesting as it suggests that the non-native English speakers from the current study outperformed the native English speakers from the original study.

In assessing the result of the current replication, it would be tempting to focus on differences between the current replication and the original study as the causal factor herein. But differences between the current study and the original study, such as the addition of the Dutch word lists, were found to be unlikely in mediating the found results as no such language bias was found during the analysis. Personal communication with the original author did also not reveal any possible weaknesses of the current replication with regard to the addition of the Dutch word lists. Another, possibly more convincing, argument in assessing the inability to replicate the results, might be the smaller sample size when compared to the original study. But this sample size was chosen based on preliminary power analysis, indicating a power level of .95 with the current sample size. The power level attained by the current study deviates negatively from this .95 level. It is possible to question these guidelines on the basis of varying effect sizes with small samples and possible overestimations in the original study. Solutions for this type of critique can range from upgrading the current OSF guidelines to running experiments based on a criterion with bigger samples, although a clear benchmark for proper power attainment is lacking. A further noteworthy finding is the lack of a correlation between respondents' English language proficiency and their attained English word performance. Such a finding might indicate that language proficiency is not a mediating factor for English word performance on the word task, although further confirmatory analysis is needed to ascertain such a claim.

[4] This was pointed out by one reviewer. This is discussed in Schoenbrodt and Perugini, 2013.

References

Beaman, C.P., Neath, I., & Suprenant, A.M. (2008). Modeling Distributions of Immediate Memory Effects: No Strategies Needed? *Journal of Experimental Psychology: LMC, 34 (1*), 219 - 239.

Champely, S. (2012). pwr: Basic functions for power analysis: Power analysis functions along the lines of Cohen (1988). R package version 1.1.1. http://cran.r-project.org/web/packages/pwr/index.html

Del Re, A. C. (2013). compute.es: Compute Effect Sizes. R package version 0.2-2. http://CRAN.R-project.org/package=compute.es

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs, *Frontiers in Psychology, 4*, 1-12.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*,*39*, 175-191.

LaPointe, L. B., & Engle, R. W. (1990). Simple and complex word spans as measures of working memory capacity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 1118–1133.

Perugini, M., Gallucci, M., & Constantini, G. (2014). Safeguard power as a protection against imprecise power estimates, *Perspectives on Psychological Science, 9* (3), 319-332.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality, 47*, 609-612.