Journal of Articles in Support of the Null Hypothesis
Vol. 15, No. 1
 1539-8714
www.jasnh.com

Implicit sound symbolism effect in lexical access, revisited: A requiem for the interference task paradigm

Chris Westbury
Department of Psychology, University of Alberta

Sound symbolism refers to a systematic association between phonemes and meaning. It has been claimed that continuant consonants are associated with round shapes, while stop consonants are strongly associated with sharp shapes. Westbury (2005) developed an implicit measure of this effect, asking participants to make lexical decisions to strings inside round or sharp frames. Decisions to all-continuant strings were faster when they were presented in compatible curvy frames and vice versa. Several unpublished attempts at replication have failed to replicate this effect. Here I re-analyze the original data and report a failure to replicate my own effect. Although the re-analysis supports the original conclusions, it also uncovers some problematic features of the original effect.

Keywords: sound symbolism, lexical access, psycholinguistics, reading, implicit interference

Correspondence should be addressed to Chris Westbury: Department of Psychology, University of Alberta, P220 Biological Sciences Building, Edmonton, AB, Canada T6G 2E9
Or: Tel: 780-492-5216, Fax: 780-492-1768, E-mail: chrisw@ualberta.ca

Acknowledgements: This work was made possible by a grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

Introduction

Sound symbolism refers to the idea that some sounds or letters have a non-arbitrary association with meaning. Although this systematic association between sound and meaning violates a basic assumption of linguistics, the arbitrariness of the sign (Hockett, 1963; Saussure, 1916/1983), there is much evidence to support the existence of sound symbolism.

One of the key sound symbolism findings that has been documented many times (e.g. Davis, 1961; Holland & Wertheimer, 1964; Maurer, Pathman, & Mondloch, 2006; Nielsen & Rendall, 2011; Ramachandran & Hubbard, 2001; Sidhu & Pexman, 2016; Westbury, Hollis, Sidhu, & Pexman, 2018) is that continuant consonants (i.e. phonemes such as /f/, /m/, and /l/ that are produced with a continuous flow of air) are more strongly associated with round shapes, while stop consonants (i.e. phonemes such as /t/ and /k/ that are produced by stopping the flow of air inside the mouth) are more strongly associated with sharp shapes. The first person to demonstrate this (following a suggestion in Usnadze, 1924) was Köhler (1929, 1947), who showed his experimental participants a round and a sharp shape and found that they strongly preferred the label *maluma* (in 1947, or *baluma* in 1929) for the round shape and the label *takete* for the sharp shape. One of the problems with Köhler's finding (and much subsequent work) is that it was limited to a forced choice decision for single pair of words, making it difficult to understand how general the effect was. Westbury (2005) tried to extend Köhler's famous finding with evidence from repeated testing, by using a paradigm that did not require participants to explicitly make decisions about meaning, but depended instead on an implicit interference effect. He asked participants to make lexical decisions about words and nonwords (NWs) that contained only stop consonants, only continuant consonants, or a mixture of both. The interference was provided by showing those letter strings inside 40 frames that were either sharp or round (see examples in Figure 1, which is discussed further below). The experimental hypothesis was that decisions to continuant strings would take longer when they were embedded in conceptually-mismatched sharp frames than when they were embedded in conceptually-matched curvy frames, and vice versa: stop consonant strings would take longer when embedded in curvy frames than in sharp frames. Westbury (2005) reported evidence of the expected effect (for NWs only), although the evidence was fragile.

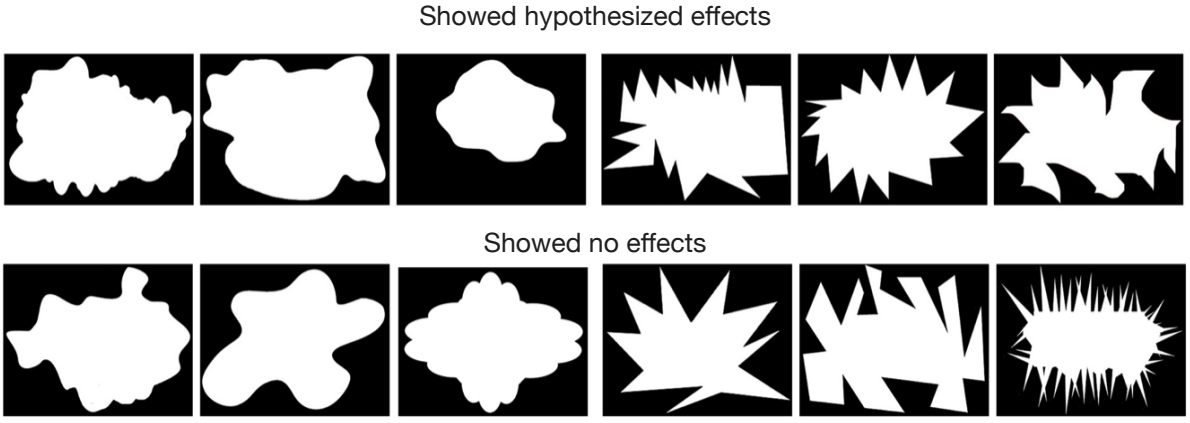

Figure 1. *Examples of frames from Westbury (2005) that either showed the expected sound symbolic interference effect for nonwords (top row) or did not that effect (bottom row). See also Figure 3.*

Analyzing the effects by participant with a 2 (curvy frame/sharp frame) × 3 (all-continuant, mixed, or all-stop strings) ANOVA, there was a marginally reliable frame × phonology interaction effect for NWs (p = 0.05). All-continuant NWs were recognized more slowly in spiky frames. Conversely, all-stop words were recognized more quickly in spiky frames than curvy frames, although the effect was marginally unreliable (p = 0.06). When the same analysis was repeated by items, the same reliable effects were found, but only when the analysis was done without the mixed strings.

Given the marginal effects and their sensitivity to the method of analysis, it is perhaps not surprising that unpublished research has failed to replicate this effect, in several different languages (personal communications). These unreported null effects are important since the original paper has been fairly widely cited (145 times by September, 2017). It is possible that that the original effects were a Type 1 error, or that some unconsidered parameters of the experiment had a systematic effect on the results. In this paper, I consider two parameters that might have impinged on the reported results: differences attributable to the individual frames and differences attributable to the size of those frames, relative to the size of the text. I also look more closely than the original paper did at the null effects reported for words. Although the original effect held up to closer scrutiny, I have replicated my colleagues in being unable to replicate the original effect.

Study 1

In the first study to be presented here, I re-analyzed the original correct response data from Westbury (2005) using linear mixed effect (LME) modelling. Since LME models can control for random effects that were irrelevant to the original hypothesis, this re-analysis constitutes a more rigorous test of the original effect. I considered random effects attributable to participants and stimulus order, and compared models using the Aikake Information Criterion (AIC, Aikake, 1973), a measure of the information loss attributable to each model, by which lower values indicate less information loss, i.e. a better fitting model.

Table 1. *Model table for LME model re-analysis of data from Westbury (2015) [Study 1], before and after removing two participants who showed an extremely large effect (see discussion in Conclusion).*

		Full data		Extreme effects removed	
Model Name	**Model description**	**AIC**	**IMPROVEMENT**	**AIC**	**IMPROVEMENT**
M0	(1 | SUBJECTID)	41428.0	BASE	37934.58	BASE
M1	(1 | ORDER)	41789.3	NO	38309.34	NO
M2	(1 | SUBJECTID) + (1 | ORDER)	41428.95	NO	37934.46	NO
M3	(1 | SUBJECTID) * (1 | ORDER)	41428.95	NO	37934.46	NO
M4	M0 + WORDNESS	41243.87	1.55E+40x	37770.59	4.07E+35x
M5	M0 + WORDNESS + PHONOTYPE	41232.01	3.60E+42x	37760.39	164x
M6	M0 + WORDNESS + PHONOTYPE + FRAMETYPE	41227.16	11.3x	37755.73	NO
M7	M0 + WORDNESS + PHONOTYPE * FRAMETYPE	41211.32	2751.77x	37741.01	16155x
M8	M0 + WORDNESS * PHONOTYPE * FRAMETYPE	41163.2	2.81E+10x	37692.36	3.67E-10x

The model comparisons are shown in Table 1. The base model M0 was defined as the model containing only random effects of participant (AIC = 41428.0). Adding a random effect of stimulus order, either as an additional predictor or in interaction with participant, did not improve the model (M1 and M2, AIC ≥ 41428.9). There were very large effects of adding in both wordness (word or nonword; M3 AIC = 41243.9) and string phonology (continuant, mixed, or stop; M4 AIC = 41232.0), each one making minimization of information loss more than 1e40 times more likely than not including it. Adding frame type (sharp/round) improved the model less markedly (M5 AIC = 41227.2, about 11.3 times more likely to minimize information). Most importantly, since it directly tests the hypothesis, including the three-way interaction between wordness, string phonology, and frame type improved the model markedly (M7 AIC = 41163.2, which suggests that this model is about 2.8e10 times more likely to minimize information loss than the model M7 that did not include the interaction).

The fitted estimates of model M7 are show in Figure 2. As originally reported, there are large differences between the strong phonology types that are mediated by frame type. In particular, NW continuant strings were correctly rejected 55 ms more quickly in round frames (Average [*SD*] estimated RT: 845.4 [120.8] ms) than NW stop strings (Average [*SD*] estimated RT: 900.9 [119.4] ms), while NW stop strings were correctly rejected 62.3 ms more quickly in sharp frames (Average [*SD*] estimated RT: 845.2 [120.8] ms) than NW continuant strings (Average [*SD*] estimated RT: 908.0 [120] ms). This hypothesis-consistent finding is muddied by the fact that mixed NWs show reaction times closer to the NW stop strings than to the NW continuant strings.

The words show a very different pattern. In the sharp frames, there are nearly identical RTs for stop words (Average [*SD*] estimated RT: 748.6 [120] ms) and continuant words (Average [*SD*] estimated RT: 743.9 [121] ms). In the round frames, there was an RT

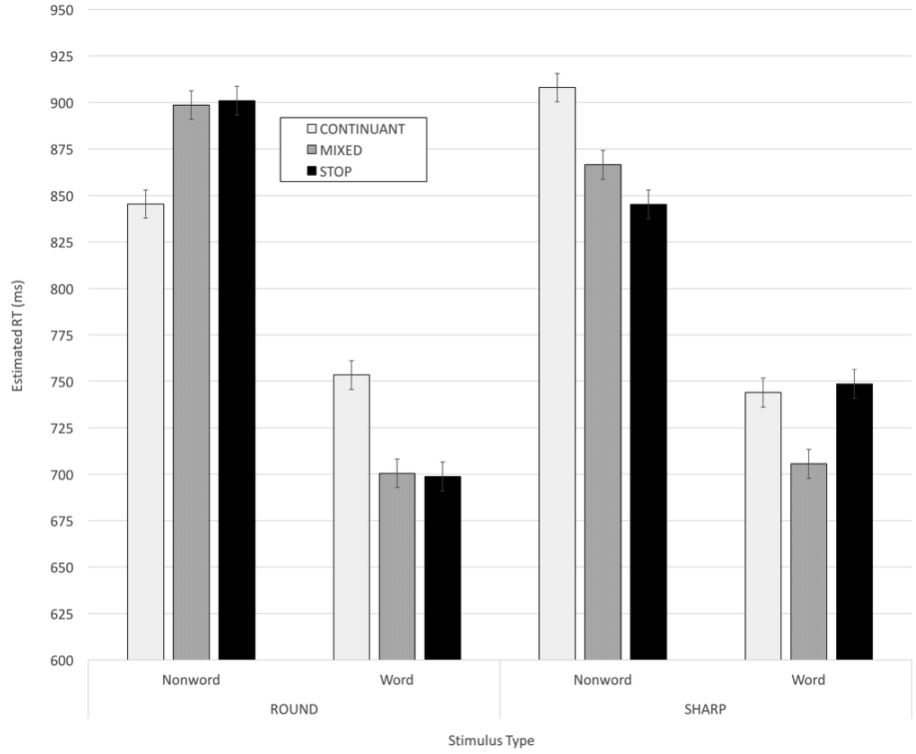

Figure 2. *Estimated RTs from re-analysis of data from Westbury (2005). Bars are SE.*

advantage of 54.6 ms for correctly accepting stop words (Average [*SD*] estimated RT: 698.8 [119] ms) over correctly accepting continuant words (Average [*SD*] estimated RT: 753.4 [121] ms).

Discussion

The results are consistent with the results originally reported in Westbury (2005), supporting the original claim of an implicit sound symbolism effect for 'pure' (all-continuant or all-stop) NW strings.

Two complications to these results (which were not reported in the original paper because I had not noticed them at the time of publication) are that i.) the 40 frames show systematically different differences between stop and continuant strings and ii.) those effects are markedly different for words and nonwords. These complications are illustrated in Figure 3, which shows the fitted estimates from the LME analysis above, by frame. Twenty-one frames showed a large (> 50 ms difference) NW continuant-stop effect in the hypothesized direction. Ten of the remaining 19 frames showed almost no effect (< 20 ms difference). It is noteworthy, since it further supports the original claim of an implicit sound symbolism effect, that 100% of the frames that showed a large effect were consistent with the predictions: i.e. all nine frames showing much faster RTs for continuant NWs were curved (Exact binomial $p = 0.0019$) and all 12 of the frames showing much faster RTs for stop NWs were sharp (Exact binomial $p = 0.00024$).

Figure 3 also shows that the words show a quite different pattern. The continuant-stop RT differences for words and nonwords are uncorrelated ($r = -0.32$, $p = 0.16$).

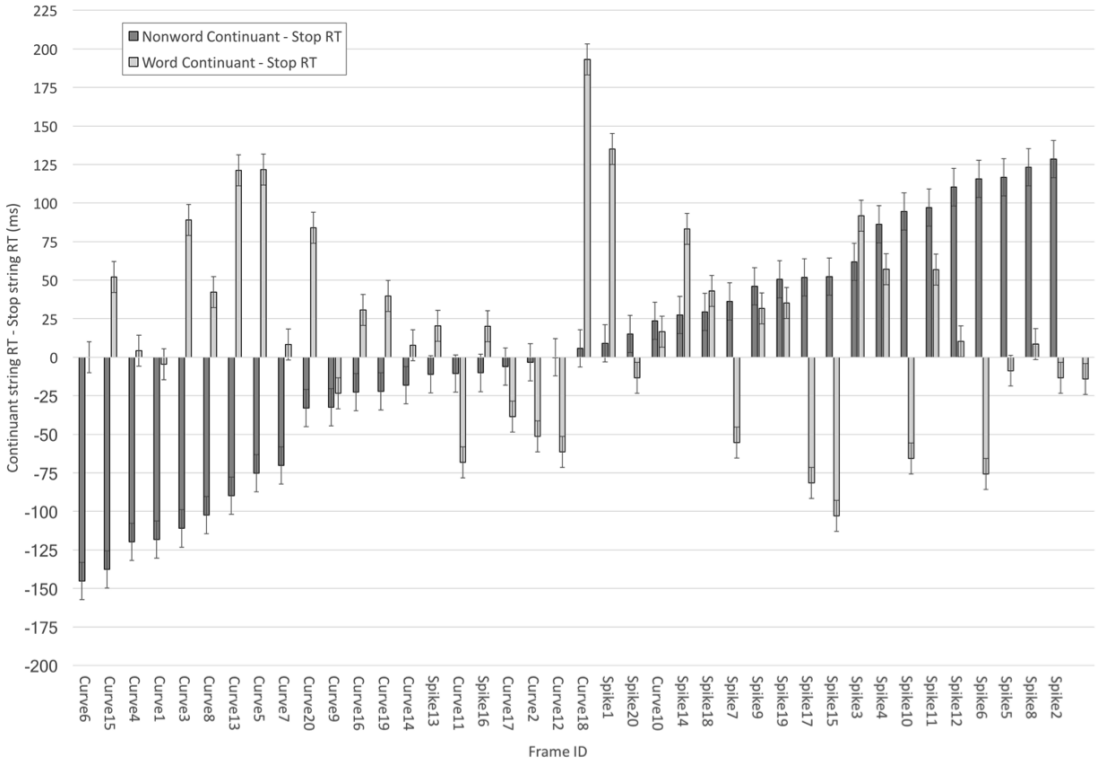

Figure 3. *Continuant string RTs – Stop string RTs, by frame, for the data from Westbury (2005), sorted by NW continuant RT – NW stop RT magnitude. Bars are SE.*

Figure 1 shows examples of frames that showed or failed to show an effect for NWs. Some of the reasons for their different effects are (perhaps) hinted at by their form. The curvy shapes that show no effects (left bottom row) all have at least one elongated arm, which possibly has the effect of making them look like "spiky curves." The sharp that failed to show an effect are less clear, although it might be speculated that some are closer to rectangular than the sharp frames that did show an effect, perhaps making their spikiness less salient as they seem more like "standard" frames. Such speculation is unnecessary since we will return to this issue when we have relevant empirical data from Study 2, to which we now turn our attention.

Study 2

A third potential problem with the implicit sound symbolism paradigm is not apparent from visual examination of the frame-shapes themselves or a re-analysis of the 2005 data. The original experiment was run on 15" iMac G3. using CRT monitors that had a 13.8" horizontal viewing area showing a resolution of 1024 x 768 pixels, for a pixel pitch of 0.46 mm. In the original experiment, the frames were 432 x 288 pixels and they were tight around the text. In this follow-up study, the size of the frames was manipulated, while keeping the text size constant, in order to test the hypothesis that framing the words tightly could have had an impact on the reported effect.

This experiment was reviewed by the Ethics Review Board at the University of Alberta.

Participants

Participants were 42 right-handed English-speaking undergraduates (16 male; 26 female) who participated in return for partial course credit. They had an average [*SD*] of 13.6 [0.74] years of education (i.e. most were in first year university). Their average [*SD*] age was 18.9 [1.08] years.

Stimuli

The same strings and forty frames that were used in Westbury (2005) were used in this experiment. However, in contrast to the original experiment, there were two frame sizes, large and small. The small frames contained white shape frames inside a rectangle of 378 x 253 pixels, centered (invisibly against a black background) inside a black rectangle of 454 x 362 pixels (11.8 cm x 9.4 cm), which subtended approximately 30 degrees of the participant's visual field. Since the background was black, the only white showing was the sharp or curvy shape inside the frame. For the large frames, the white shape frame was enlarged close to the edge of the larger 454 x 362 pixel rectangle (see examples in Figure 4).

Method

Stimuli were presented using ACTUATE software (Westbury, 2007) running under Apple's OS 10.6 on G4 Mac Minis connected to 17.1" (15.1" x 15.0"; 1280 x 1024 pixels) Samsung SyncMaster 713V monitors, with a pixel pitch of 0.26 mm. The experiment was run in one of three testing rooms constructed to reduce outside noise. participants were seated

approximately 45 cm from the screen. They were shown written instructions that were simultaneously presented verbally by a research assistant. The instructions asked them to decide as quickly and accurately as they were able if each string was an English word, indicating their choice by pressing the 'x' key (for "wrong") or the 'c' key (for "correct"). As in the original 2005 experiment, neither the written instructions nor the experimenter made any mention of the frame manipulation (i.e. participants were not informed that the strings would be appearing in different frame shapes).

Strings were presented within the framed shapes in 90 point Times font. Each string was preceded by a '+' to orient the participants to the next stimulus. This cue was presented for a random amount of time uniformly sampled between 0 and 1500 ms. The ISI was 1000 ms.

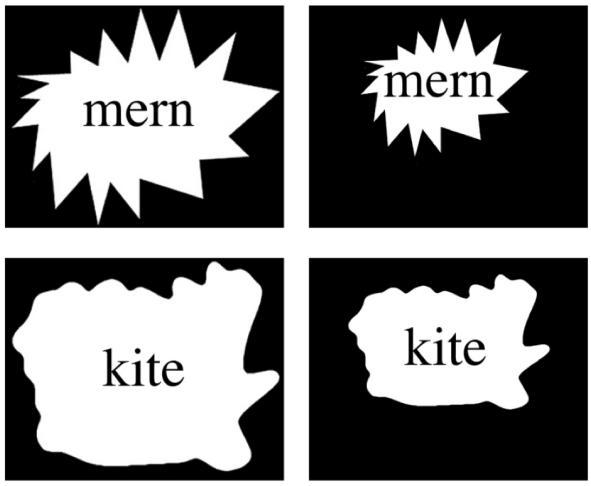

Figure 4. *Examples of letter stimuli in large (left) or small (right) frames.*

Each participant began with five practice trials to familiarize them with the procedure. These trials were discarded before data analysis. They then made 120 decisions, 60 in large and 60 in small frames, divided equally between word and NW decisions and between all-continuant, all-stop, and mixed strings. Each participant's input list was randomly generated individually, so that all strings would be likely to be encountered equally often in all frame size x frame shape conditions.

After discarding the 352 erroneous responses, data were trimmed by first removing 23 responses < 400 ms and all eight responses > 4000 ms, and then removing any remaining responses that fell outside of +/− 3z (Average [*SD*] correct RT: 781 [363] ms; 103 responses were removed).

Results

The results were again analyzed using linear mixed effects modeling. The models are summarized in Table 2. The key findings are that the reliable three-way interaction between wordness (word/nonword), string phonology(stop/continuant/mixed), and frame type (sharp/curvy) was replicated (M6, AIC = 56866, 1.7e12 times more likely to minimize

Table 2. *Model table for LME model analysis of data from Study 2, with all participants (left) and with only the slowest 50% of participants (right; see discussion in Conclusion)*

		Full data		Fast participants removed	
Model name	**Model Description**	**AIC**	**IMPROVEMENT**	**AIC**	**IMPROVEMENT**
M0	(1 | SUBJECTID)	57234.1	BASE	29718.27	BASE
M1	(1 | ORDER)	57937.0	NO	29794.39	NO
M2	(1 | SUBJECTID) + (1 | ORDER)	57236.1	NO	29720.27	NO
M3	M0 + WORDNESS	56935.0	8.8E64x	29478.02	1.48E52x
M4	M0 + WORDNESS + PHONOTYPE	56926.0	91x	29466.75	280x
M5	M0 + WORDNESS + PHONOTYPE + FRAMETYPE	56922.1	NO	29462.27	NO
M6	M0 + WORDNESS * PHONOTYPE * FRAMETYPE	56865.7	1.7E12x	29407.55	7.16E12x
M7	M0 + WORDNESS * PHONOTYPE * FRAMETYPE * SIZE	56783.1	8.6E17x	29307.4	5.59E21x

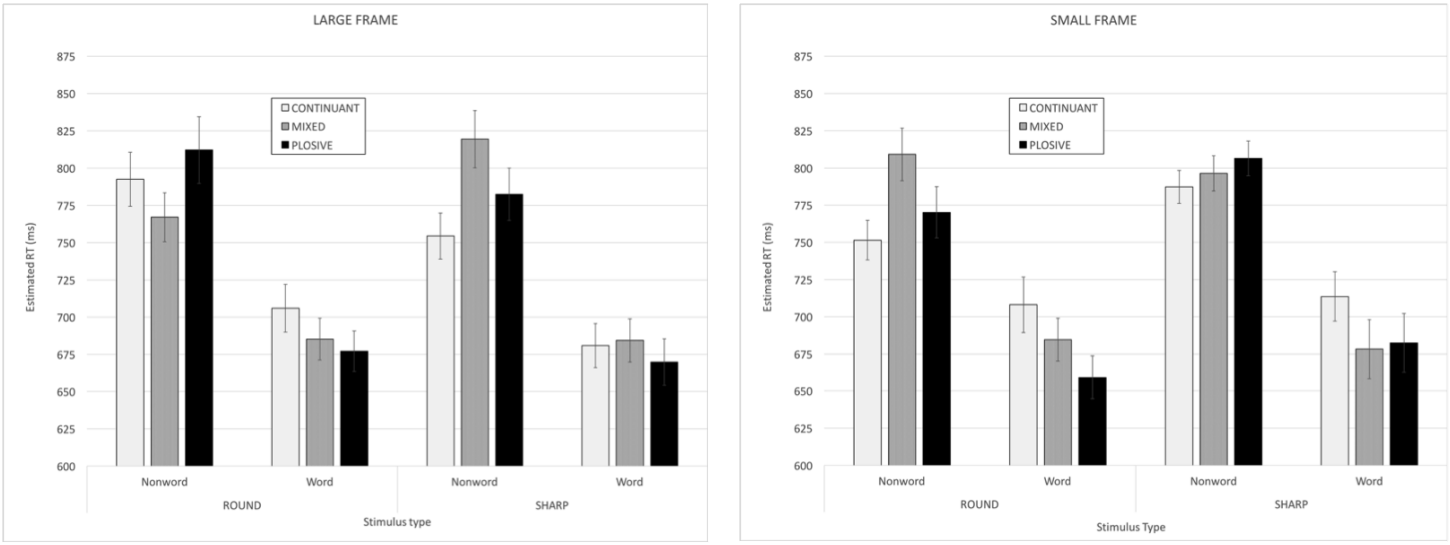

Figure 5. *LME model estimated RTs from Study 2. Bars are SE.*

information loss than the same model without interactions, M5, AIC = 56922), and that there was a reliable four-way interaction when frame size was added to that three-way interaction (AIC = 56783, 8.6e17 times more likely to minimize information loss than model M5).

These results are graphed in Figure 5. Although the three-way interaction was reliable, the key hypothesized directional difference between nonword phonology and frame type was not replicated. Nonwords containing continuants were recognized more quickly than nonwords containing stops when they were presented in both sharp and round frames of either size (Large frame continuant-stop differences: −19.6 ms for round frames and −28.0 ms for sharp frames; Small frame continuant-stop differences: −18.7 ms for round frames and −19.1 ms for sharp frames).

The effects of the frame size manipulation are shown in Figure 6. The largest effects are that RTs to phonologically pure (all-continuant or all-stop) NWs are much (> 40 ms) slower in large frames than in small frames. This may simply reflect the fact that the location of the strings is better picked out by the small frames, although this does not explain the fact that *mixed* NW strings are not recognized slower (but rather > 40 ms faster) in large frames than in small frames.

Figure 7 shows the correlation between the difference between the continuant NW RTs and the stop NW RTs in the two studies, by frame, to test whether the frame effects shown in Figure 3 are consistent. They are not. The correlations between the effects sizes by frame are close to zero for both the small frames and the large frames. Moreover, there was no clean split between the curved and sharp frames by the size of the continuant NW RT and the stop NW RT differences. For the small frames, four of the ten frames with a continuant-stop difference < −50 ms (i.e. faster responses for continuant strings) were sharp (Exact binomial $p = 0.20$). Only two frames had a difference > 50 ms (i.e. faster responses for stop strings) and, contrary to the original hypothesis, they were both curvy. Similar statistically unreliable effects were found for the large frames.

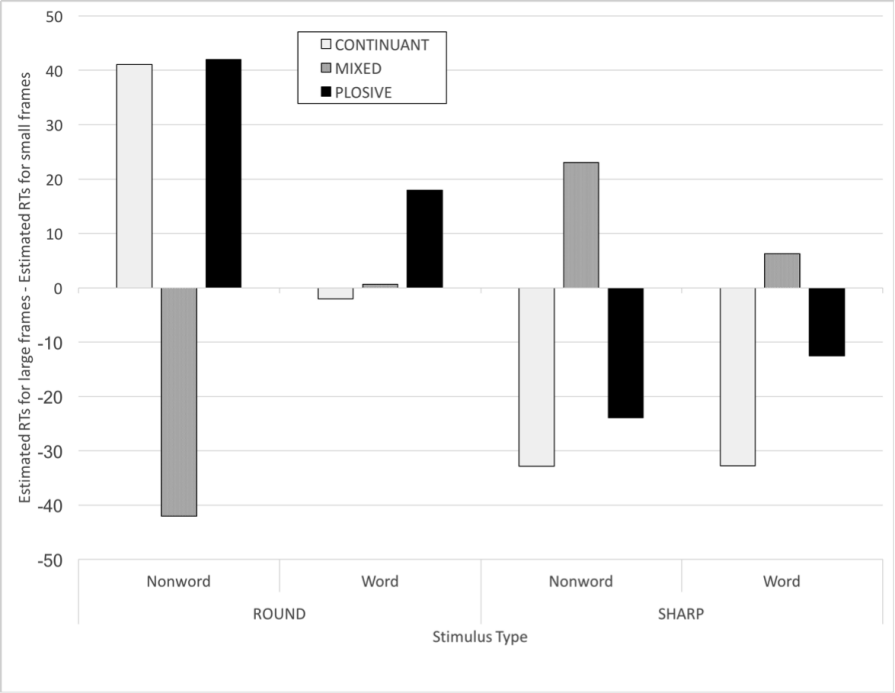

Figure 6. *Differences between LME-estimated RTs for large and small frames*

General Discussion

The effects published in Westbury (2005) stood up to further analysis in Study 1. LME modelling found strongly reliable effects of the same kind as were originally reported. Moreover, an analysis not reported in the original paper found that the observed distribution of frame by effects (i.e. a large continuant string advantage only for curvy frames and a large stop string advantage only for sharp frames) is, by exact binomial probability, very unlikely to have occurred by chance.

Nevertheless, an attempt at replicating and possibly explaining previous failures to replicate also failed to find support for the original hypotheses. Although there was a similar three-way interaction between frame type, phonological category, and wordness, the direction of the nature was inconsistent both with the original hypotheses and with the results from the original 2005 experiment.

There are a few implications to be drawn from these results.

One is that hand-drawn frames may be insufficient for the purposes of studying sharp/curvy interference. As suggested by Figure 1, some curvy frames are straighter than other curvy frames and sharp frames with small 'teeth' around a curvy shape can seem curvy. It may be better either to use mathematical methods to control sharpness/curviness systematically (as in, e.g. Nielsen & Rendall, 2011) or to make curvy and sharp frames that are more closely matched to each other by rounding the teeth of a sharp frame to turn it into a curvy frame, so that the general shape is held constant between conditions (as in, e.g., Sidhu & Pexman, 2015). Moreover, since there seem to be large frame effects, any future work dependent on frame manipulations should either follow Westbury (2005) in using many different frames to average out the effects of individual frames, or undertake a preliminary study to identify the frames most reliably associated with the effects of interest.

The other implication, suggested by the results from Study 2, is that there are

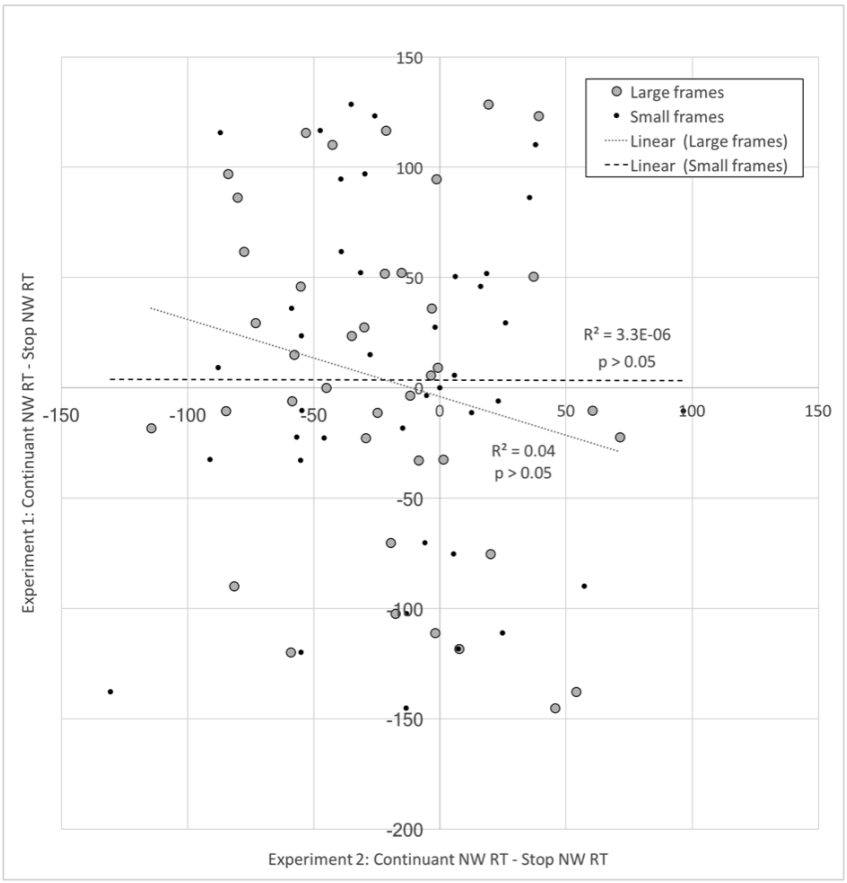

Figure 7. *Correlation of continuant NW − stop word RT difference, between the original experiment (Westbury, 2005), on the y axis, and the results from Study 2 for large and small frames, on the x-axis.*

RT effects attributable to the size of the frame, though there is no evidence that tight frames (which were consistent with the original 2005 experiment) show the hypothesized implicit sound symbolism effect, while looser frames do not. As shown in Figure 6, the effects attributable to frame size are not only large (as large as 40 ms for all NW types in round frames) but also variable. For example, small frames are associated with faster RTs for all-continuant and all-plosive NW decisions when they are round, but the direction of this effect is reversed when the frames are sharp. These effects of frame size and shape, independent of whether the string was composed of continuant or plosive consonants, are simply a nuisance with respect to the hypotheses being tested, and make it more difficult to interpret any results.

Although I have reported the results of only a single new experiment here, that experiment was motivated by repeated personal communications of a failure to replicate the original effect. In the context of those motivating reports, the bulk of the evidence now suggests that the implicit sound symbolism effect is weak, non-existent, or controlled by some factor. One possibility for that factor is subject effects. It seems unlikely that there could be systematic differences between the participant groups, who were very similar in their demographic makeup and drawn from the same participant pool (albeit several years apart). However, it is possible that the effects were dominated by a few participants. Following standard practice, both experiments trimmed the data by removing RT outliers globally, i.e. without regard to the condition in which they were seen. After data trimming,

the original experiment had two participants who showed NW curve-spike differences, in the predicted direction, of over 400 ms (>= 4.9z from the average observed difference in that experiment). The largest difference in the second experiment (also in the predicted direction) was 263 ms. I removed the two participants who showed the very extreme effects in the original experiment, and repeated the LME analysis above. The pattern of model analysis (shown in Table 1) was very similar to the original analysis, and the results closely resembled those shown in Figure 2, with the same effects replicated. The two extreme values were not driving the effect.

It is nevertheless notable that the RTs in the original experiment were slower and more variable than most lexical decision RTs (Average [*SD*]: 725 [289] ms for words; 877 [348] ms for NWs, as compared to Experiment 2: Average [*SD*]: 684 [196] ms for words; 787 [243] ms for NWs). This suggests that the original participants may have been unusual in some way. It is possible that the reported effect is seen for slower but not quicker readers, perhaps because slower readers are exposed to the frames for longer. To test this, I eliminated the 50% of the participants who had the quickest average RTs, collapsed across words and NWs, in Experiment 2 and re-analyzed the data with the slower participants only. In this subset of 21 people, the average [*SD*] RTs were as slow or slower than the RTs in the original Westbury (2005) experiment (737 [199] ms for words in the large frames; 733 [210] ms for words in the small frames; 893 [271] ms for NWs in the large frames; and 888 [260] ms for NWs in the small frames). The pattern of model analysis (Table 2) was very similar to the original analysis of the full dataset. Moreover, the key result from the replication experiment reported above was unchanged: Nonwords containing continuants were recognized more quickly than nonwords containing stops when they were presented in sharp and round frames of either size. This suggests that the original effect cannot be explained simply by the fact that it was obtained with slow readers.

Unfortunately, it remains unclear exactly why there was an implicit interference effect in Westbury (2005). However, the fact that is has now failed to replicate multiple times suggests that it was probably a Type 1 error. The interference task paradigm appears to be an unsuitable paradigm for studying implicit sound symbolism.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716-723.

Davis, R. (1961). The fitness of names to drawings: A cross-cultural study in Tanganyika. *British Journal of Psychology, 52*, 259-268.

Hockett, C. (1963). The problem of universals in language. In J. Greenberg (Ed.), *Universals of language*. Cambridge, MA: MIT Press.

Holland, M. K., & Wertheimer, M. (1964). Some physiognomic aspects of naming, or, *maluma* and *takete* revisited. *Perceptual and Motor Skills, 19*, 111-117.

Köhler, W. (1929). *Gestalt psychology*. New York, NY: Liveright.

Köhler, W. (1947). *Gestalt psychology* (2nd ed.). New York, NY: Liveright.

Maurer, D., Pathman, T., & Mondloch, C. J. (2006). The shape of boubas: Sound–shape correspondences in toddlers and adults. *Developmental Science, 9*(3), 316-322.

Nielsen, A., & Rendall, D. (2011). The sound of round: Evaluating the sound-symbolic role of consonants in the classic Takete-Maluma phenomenon. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 65*(2), 115.

Ramachandran, V. S., & Hubbard, E. M. (2001). Synaesthesia—A window into perception, thought and language. *Journal of Consciousness Studies, 8*, 3-34.

Saussure, F. (1916/1983) *Course in general linguistics*. Eds. C. Bally & A. Sechehaye. Trans. R. Harris. La Salle, IL: Open Court.

Sidhu, D. M., & Pexman, P. M. (2015). What's in a name? sound symbolism and gender in first names. *PloS one, 10*(5), e0126809.

Sidhu, D. M., Pexman, P. M., & Saint-Aubin, J. (2016). From the Bob/Kirk effect to the Benoit/Éric effect: Testing the mechanism of name sound symbolism in two languages. *Acta Psychologica, 169*, 88-99.

Usnadze, D. (1924). Ein experimenteller Beitrag zum Problem der psychologischen Grundlagen der Namengebung. *Psychological Research, 5*(1), 24-43.

Westbury, C. (2005). Implicit sound symbolism in lexical access: Evidence from an interference task. *Brain & Language, 93*(1), 10-19.

Westbury, C. (2007). *ACTUATE: Assessing cases: The University of Alberta Testing Environment*. Retrieved from http://www.psych.ualberta.ca/~westburylab/downloads/actuate.download.html

Westbury, C., Hollis, G., Sidhu, D. M., & Pexman, P. M. (2018). Weighing up the evidence for sound symbolism: Distributional properties predict cue strength. *Journal of Memory and Language, 99*, 122-150.

Received: 7.7.2017
Revised: 9.28.2017
Accepted: 9.29.2017