


Cognitive Psychology

Conducting Language Production Research Online: A Web-based Study of Semantic Context and Name Agreement Effects in Multi-Word Production

Jieying He¹ ^a, Antje S. Meyer², Ava Creemers³, Laurel Brehm³

¹ Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands; International Max Planck Research School for Language Sciences, Nijmegen, The Netherlands, ² Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands; Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands, ³ Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

Keywords: bayesian analysis, name agreement, semantic context, picture naming, web-based experiments

<https://doi.org/10.1525/collabra.29935>

Collabra: Psychology

Vol. 7, Issue 1, 2021

Few web-based experiments have explored spoken language production, perhaps due to concerns of data quality, especially for measuring onset latencies. The present study highlights how speech production research can be done outside of the laboratory by measuring utterance durations and speech fluency in a multiple-object naming task when examining two effects related to lexical selection: semantic context and name agreement. A web-based modified blocked-cyclic naming paradigm was created, in which participants named a total of sixteen simultaneously presented pictures on each trial. The pictures were either four tokens from the same semantic category (homogeneous context), or four tokens from different semantic categories (heterogeneous context). Name agreement of the pictures was varied orthogonally (high, low). In addition to onset latency, five dependent variables were measured to index naming performance: accuracy, utterance duration, total pause time, the number of chunks (word groups pronounced without intervening pauses), and first chunk length. Bayesian analyses showed effects of semantic context and name agreement for some of the dependent measures, but no interaction. We discuss the methodological implications of the current study and make best practice recommendations for spoken language production research in an online environment.

Introduction

The use of internet-based experiments for behavioral research has gained in popularity over the last few years, driven by the increasing ease and efficiency with which larger and more diverse samples of participants can be reached (e.g., Reimers & Stewart, 2015) and by the Covid-19 pandemic (e.g., Sauter et al., 2020). In psycholinguistics, web-based variants of sentence comprehension and word recognition experiments elicit good quality data in questionnaires or typed responses (e.g., Cooke et al., 2011; Schnoebelen & Kuperman, 2010). However, web-based experiments of spoken production are still uncommon. At the time of planning this study, there were two main concerns: one concerned the quality of speech recording made outside of a laboratory environment, the other concerned the precision of measurement of speech onset latencies due to potentially poor audiovisual synchrony. That is, it was not clear whether the timing of visual stimuli on the participant's screen and of the onset of the recording of their responses could be controlled precisely enough to obtain use-

ful measures of speech onset latencies (see also Bridges et al., 2020). The current study therefore explored the usefulness of dependent measures that did not depend on this synchrony, but were derived from the durations and fluency of the participants' utterances. Meanwhile, recent speech production studies have shown that onset latencies can in fact be measured with good accuracy in web-based platforms (e.g., Fairs & Strijkers, 2021; Stark et al., 2021; Vogt et al., 2021). We review these studies in the Discussion section.

The present study measured utterance durations and utterance internal pauses (indexing speech fluency) offline during multiple-utterance production. Unlike speech onset latency, the precision of temporal characteristics within participants' audio recordings can be guaranteed sufficiently in web-based experiments: the interval between the recorded utterance onset and offset (i.e., utterance duration), or the interval between the offset of the first word and the onset of the second word (i.e., pause time) can be measured from the recording itself. These measures are limited by the quality of the participants' recording equipment and

^a Correspondence concerning this article should be addressed to Jieying He, Max Planck Institute for Psycholinguistics, P.O. Box 310, 6500 AH Nijmegen, The Netherlands. Email: Jieying.He@mpi.nl.

the researcher's speech analysis tools, but not by the issue of audiovisual synchrony, which means that regardless of how successful the audiovisual synchrony is, we should be able to obtain reliable measurements.

Language production work has typically exploited speech onset latency as the dependent variable, but variations in other characteristics of the utterance, such as utterance duration and speech fluency (indexed by pauses), are also promising measures for examining multi-word production (e.g., Ferreira & Swets, 2017; Kandel et al., 2021; Momma & Ferreira, 2019). This is because speakers do not necessarily fully plan multi-word utterances before beginning to speak, but rather often continue planning while articulating their utterance. The clearest evidence for this comes from studies recording participant's eye movements while they are describing scenes or events (e.g., Griffin & Bock, 2000; Konopka, 2019). Most relevant to the present study are multiple-object naming studies (e.g., Belke & Meyer, 2007; Meyer et al., 2012; Mortensen et al., 2008), which have showed that when speakers are asked to name sets of three or more objects, they usually fixate upon them in the order of mention, with the eyes running slightly ahead of the articulation of the object names. Speakers typically initiate their utterance after the shift of gaze to the second or third object. This pattern shows that speech planning continues after utterance onset. Since an upcoming word may be planned while another word is being articulated, the difficulty of word planning may be reflected in the time elapsed between word onsets, where speakers may either stretch words or insert pauses between them. Consequently, variation in the difficulty of planning processes can manifest itself not only in onset latencies, but also in utterance durations and speech fluency (see also E.-K. Lee et al., 2013).

To investigate how speech production research can be done outside of the laboratory by measuring utterance durations and speech fluency, we created a modified blocked-cyclic naming paradigm to examine two previously studied phenomena related to lexical selection: semantic context and name agreement effects. The design of the modified blocked-cyclic paradigm was inspired by work of Belke and Meyer (2007), who explored semantic context effects in picture naming. The semantic context effect is the finding that it is more difficult to name multiple objects from the same semantic category (a homogeneous context) than from different semantic categories (a heterogeneous context). In most semantic context experiments, one picture is presented per trial and onset latencies are measured (e.g., Damian et al., 2001; Damian & Als, 2005). However, Belke and Meyer (2007, Experiment 1b), explored semantic context effects during multiple object naming in young (college-aged) and older (52–68 years) speakers. On each trial four objects belonging to the same or different semantic categories were presented simultaneously on the screen and had to be named. The authors found small but significant semantic context effects on word durations for both groups of speakers, and a significant semantic context effect on pause rate for the older speakers. This indicates that semantic context effects can be obtained on measures such as utterance durations and speech fluency. These measures should remain reliable in web-based research because they

are derived from the participants' speech alone rather than the timing of their speech relative to a stimulus.

The paradigm used in the current study was further inspired by studies on rapid automatized naming (RAN), used primarily in neuropsychological work. In a RAN task, a set of familiar items (e.g., five objects or digits) repeated multiple times across rows of a grid is named as quickly as possible, and the total naming time of the grid is measured (Denckla & Rudel, 1976). There are large individual differences in total naming times. Moreover, total naming times depend also on properties of the materials such as the word frequency and phonological neighborhood density of the object names (Araújo et al., 2020). This implies that when objects are repeatedly named in a grid, variation in the difficulty of speech production can be reflected in total naming times.

Inspired by these two lines of work, we created a modified blocked-cyclic naming paradigm suitable for web-based research. On each trial, participants were asked to name sixteen pictures that were simultaneously presented in a 4 × 4 grid. Each set of sixteen pictures consisted of repetitions of four pictures which belonged either to the same semantic category or to different semantic categories, quadrupling the number of pictures named per trial in Belke and Meyer (2007). Orthogonally, name agreement for the pictures was varied. We measured five main dependent variables: accuracy, utterance duration, total pause time, total chunk number, and first chunk length. A chunk was defined as a group for words produced without intervening pause longer than 200 ms (for details, see Methods). While we were not entirely confident about the reliability of onset latencies, we also measured them, allowing us to make a rough comparison with lab-based studies.

The modified blocked-cyclic naming paradigm was used to examine whether effects of semantic context and name agreement would be obtained on dependent variables that can be measured reliably on web-based experimental platforms. We selected these independent variables because they were deemed likely to affect lexical selection in different ways. As noted earlier, the semantic context effect is the finding that speakers are slower and less accurate to repeatedly name small sets of objects in homogeneous contexts than in heterogeneous contexts (e.g., Belke & Meyer, 2007; Damian & Als, 2005; Damian et al., 2001). The semantic context effect has been attributed to the selection of lexical-semantic entries (i.e., lemmas): selecting a target lexical representation is more difficult in the context of semantically related than unrelated items (Damian et al., 2001). Importantly, the semantic context effect takes some time to build up: Typically, participants show either no semantic interference effect or a semantic facilitation effect when they name the pictures for the first time, but from the second cycle onward, they display a stable semantic interference effect (Belke, 2017; Belke et al., 2005; Damian & Als, 2005). Given that semantic context effects were mainly found on word durations in multiple object naming (Belke & Meyer, 2007), we predicted that in our paradigm semantic context effects would start to emerge, especially on the measure of utterance durations, when participants began to name the second row of objects.

Name agreement is the extent to which participants

agree on the name of a picture. The name agreement effect refers to the finding that naming a picture with high name agreement (e.g., a picture of a *banana*) is faster and more accurate than naming a picture with low name agreement (e.g., a picture of a piece of furniture which could be called *sofa*, *couch*, or *settee*; Alario et al., 2004; Vitkovitch & Tyrrell, 1995). Name agreement effects come from multiple sources. The name agreement effect is found for objects that are often incorrectly named (e.g., *celery*, which is commonly misidentified as *rhubarb*, *Chinese leaves*, or *cabbage*), reflecting difficulty in object recognition. The effect has also been obtained for objects with multiple plausible names (e.g., a *jumper* is also called *sweater*, *pullover*, *jersey*, or *sweatshirt*), reflecting difficulty at the lexical selection stage of spoken language production (Alario et al., 2004; Shao et al., 2014; Vitkovitch & Tyrrell, 1995). The present study focused on the latter effect: these low name agreement pictures evoke more lexical candidates than pictures with high name agreement, and hence, it takes longer to eliminate candidates and select one name. Thus, we predicted that name agreement would affect utterance durations, total pause time and chunk measures.

The effects of semantic context and name agreement are interesting to investigate in tandem because their relationship can provide some insight into how lexical selection is achieved in speech production. Existing models proposed to account for semantic context effects (e.g., Abdel Rahman & Melinger, 2009; Howard et al., 2006; Oppenheim et al., 2010) disagree on whether lexical selection during spoken language production is competitive or not. This disagreement means that these models make different predictions about whether increasing the number of activated lemmas during lexical selection will increase semantic context effects. Models with lexical competition (e.g., Abdel Rahman & Melinger, 2009; Howard et al., 2006) predict that semantic context should interact with name agreement, such that the semantic context effects would be stronger for low name agreement pictures than high name agreement pictures. By contrast, models not assuming lexical competition predict that semantic context effects should not be influenced by name agreement (e.g., Oppenheim et al., 2010).

Methods

Participants

We recruited 41 native Dutch speakers (36 females, $M_{\text{age}} = 22$ years, range: 19 - 26 years) from the participant pool at the Max Planck Institute for Psycholinguistics. This is about twice the sample size used in most semantic context experiments (e.g., 16 participants in Belke & Meyer, 2007; 24 participants in Damian & Als, 2005) and seemed appropriate for an exploratory study. All participants reported normal or corrected-to-normal vision and no speech or hearing problems. They signed an online informed consent form and received a payment of €6 for their participation. The study

was approved by the ethics board of the Faculty of Social Sciences of Radboud University.

Apparatus

The experiment was implemented on FRINEX (FRamework for INteractive EXperiments), a web-based platform developed by the technical group at the Max Planck Institute for Psycholinguistics (for details, see Withers, 2017). It was displayed on the participants' own laptops; we restricted participation to 14- or 15.6-inch laptops with Google Chrome, Firefox, or Safari web browsers. Participants' speech was recorded by a built-in voice recorder of the web browser. WebMAUS Basic was used for phonetic segmentation and transcription (<https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/WebMAUSBasic>). Praat (Boersma & Weenink, 2009) was then used to extract the onsets and offsets of all segmented responses.

Materials

Thirty-two pictures with one- or two-syllable primary names (see Appendix A, Table A1) were selected from the MultiPic database of 750 single-object drawings (Duñabeitia et al., 2018), which provides language norms (e.g., name agreement, visual complexity) in standard Dutch. Of these, sixteen were high name agreement pictures, all with name agreement percentage of 100%, and sixteen were low name agreement pictures, with name agreement percentages between 50% and 85% ($M = 64\%$, $SD = 11\%$). Independent *t*-tests revealed that the two sets of pictures differed significantly in name agreement, but not in any of ten other psycholinguistic attributes¹. For all low name agreement pictures, their first and second modal names in the MultiPic database share the same semantic features (e.g., *kat* 'cat' and *poes* 'cat'), as judged by a native speaker of Dutch.

The sixteen high name agreement and sixteen low name agreement pictures were selected from four semantic categories (animal, body part, clothing, and tool), with four of each semantic category. Each set of sixteen pictures was used to make a matrix of 4×4 picture grids such that the rows corresponded to the categories and thus formed homogeneous stimulus sets of four pictures each, whereas columns formed the sets for heterogeneous condition of the same size. Two picture names in each row and in each column were monosyllabic and two were bisyllabic. Pictures were selected to minimize within-category visual similarity and avoid shared initial phoneme or letter.

To equate the semantic similarity between the high and low name agreement conditions, we calculated the semantic similarity of all six pairs within each four-picture set by using *sub2vec* (van Paridon & Thompson, 2021). In homogeneous sets, semantic similarities of the pairwise combi-

¹ Ten matched variables: visual complexity, age-of-acquisition, word frequency, number of syllables, number of phonemes, word prevalence, phonological neighborhood frequency, phonological neighborhood size, orthographic neighborhood frequency, and orthographic neighborhood size.

nations of all pictures per set were matched across semantic categories by name agreement. Independent sample *t*-test showed that there was no difference in semantic similarity between high and low name agreement pictures ($t_{(46)} = 0.004$, $p = 0.997$). In heterogeneous sets, semantic similarities for the pairwise combinations of all pictures per set were also matched. Independent *t*-test also showed that there was no difference on semantic similarities between high and low name agreement pictures in each heterogeneous set ($ts < -0.6$, $ps > 0.01$).

On each trial, a 4×4 picture grid was presented from the matrix described above. There were eight picture grids (four for homogeneous trials, four for heterogeneous trials) for each name agreement condition, resulting in sixteen picture grids in total (i.e., 16 trials). Each picture grid was shown three times in different test blocks, which results in 48 trials in total. This means each individual picture was repeated six times (twice per block: once for a homogeneous picture grid, and once for a heterogeneous picture grid) during the experiment. Sixteen additional pictures (combined into four picture grids) were selected from the same database as practice stimuli, resulting in four practice trials.

Design

Semantic context (homogeneous, heterogeneous) and name agreement (high, low) were both treated as within-participant variables; both were randomized within experimental blocks and counterbalanced across participants. The same four pictures per homogeneous or heterogeneous set were presented in a different arrangement across blocks and participants with a Latin square design so that each item appeared in each ordinal position. Within a picture grid, note that the same items did always follow each other (e.g., *leeuw* 'lion' always followed *muis* 'mouse'). A unique order of displays was created for each participant with the Mix program (van Casteren & Davis, 2006), with the constraints that homogeneous and heterogeneous trials alternated, trials from the same semantic category were not presented consecutively, and the last picture on a trial was not the same as the first picture on the next trial.

Procedure

Participants were tested on the web² with the instructions that they should perform this experiment in a quiet room with the door shut and with potentially distracting electronic equipment turned off. They were told to imagine that they were in a laboratory during the experiment. We asked for permission to record before the test began. At the beginning of the test, participants were asked to familiarize themselves with all pictures and name them quickly in Dutch. Familiarization trials began with a fixation cross presented for 500 ms, followed by a blank screen for 500 ms. Then, a picture appeared on the screen for a 2-second period during which participants were asked to name the pic-

ture in Dutch as quickly and accurately as possible. Finally, a blank screen was presented for 1500 ms before the start of the next trial.

A practice session of four trials was followed by the three blocks of experimental trials. Participants took a short break after each block of sixteen trials. The whole experiment lasted 30 minutes. Practice and experimental trials began with a fixation cross presented for 500 ms, followed by a blank screen for 500 ms. Then a 4×4 picture grid appeared on the screen in which sixteen pictures were presented simultaneously for up to 30 seconds. Participants named the sixteen pictures one by one in order from left to right starting with the first row as quickly and accurately as possible. They ended the trial by a mouse click. If they had not finished within 30 seconds, the picture grid disappeared automatically. A blank screen was presented for 1500 ms before the onset of the next trial. An example of a trial is shown in [Figure 1](#).

Analysis

Five main dependent variables were coded to index naming performance. Production *accuracy* indexes the proportion of trials where all sixteen pictures were named correctly. Participants were not presented with the expected names of the pictures in the familiarization stage, as it was impossible to give them timely feedback on their naming responses and we did not want to ask them to use picture names they would not spontaneously use. Therefore, we later coded any reasonable naming responses as correct. Picture names were coded as correct if they matched any of the multiple names given to the picture in the MultiPic database (Duñabeitia et al., 2018); if they were diminutive versions of the multiple names (e.g., *big* 'piglet' was named as *biggetje* 'little piglet'), or if they were judged reasonable by trained research assistants (e.g., *gier* 'vulture' was named as *havik* 'hawk').

For trials where all pictures were named sensibly and without hesitations or self-corrections (hereafter, "fully correct trials"), we calculated two main time measures. *Utterance duration* was defined as the time interval between the utterance onset of the first picture name and the utterance offset of the sixteenth picture name. This reflects how long participants took to produce all sixteen picture names. *Total pause time* was defined as the sum of all pauses between picture names. This reflects the planning done between producing responses.

For these fully correct trials, we also examined how participants chunked or grouped their sixteen responses. Since earlier studies of spontaneous speech coded silent durations longer than 200 ms as silent pauses (e.g., Heldner & Edlund, 2010), we coded the responses that occurred with 200 ms or less between them as a single response chunk. *Total chunk number* refers to how many response chunks participants made on one trial, with a larger number of total

² Here is an example of the experiment for one participant: https://frinexproduction.mpi.nl/image_naming_experiment/?stimulus-List=List1

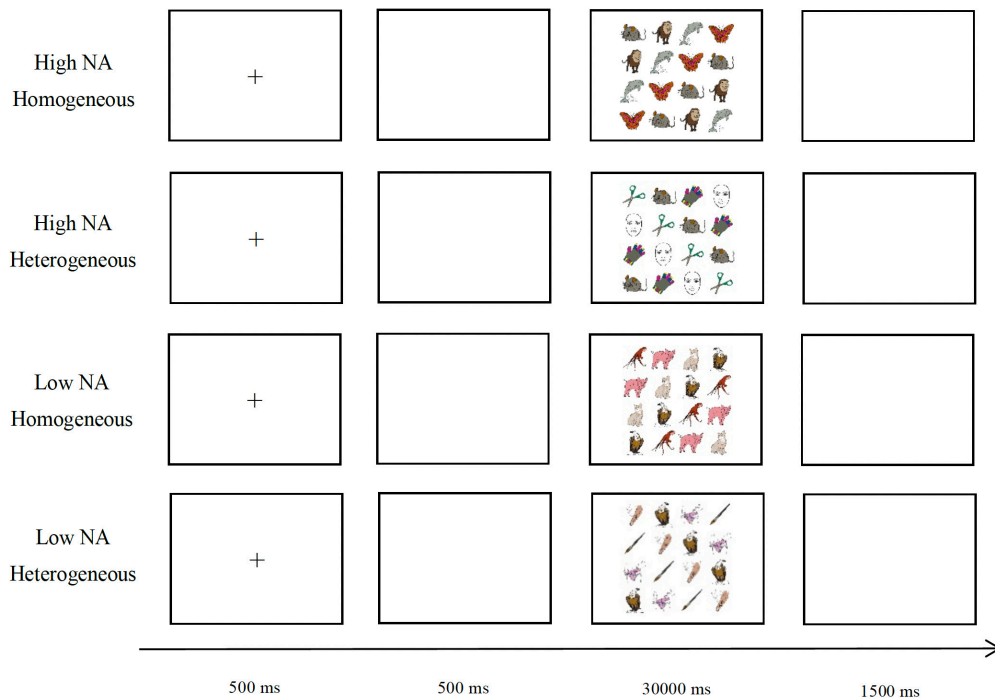


Figure 1. Trial examples of four conditions

NA refers to name agreement. Heterogeneous sets include one picture from each homogeneous set.

response chunks meaning more separate planning units for production. *First chunk length* refers to how many names participants produced in their initial response, and provides a measure of how much information participants planned before starting to speak. In addition to the five primary measures of interest, we also measured *onset latency*, defined as the interval from the onset of stimulus presentation to the onset of the utterance, which indexes the beginning stages of speech planning.

Bayesian mixed-effect models were conducted to assess the likely magnitude of the effects and quantify the size of parameters and the uncertainty around them (Nicenboim & Vasishth, 2016). Bayes factors were computed to evaluate the evidence in favor of or against the effects. For these analyses, we used R version 4.0.3 (R Core Team, 2020) with the package *brms* (version 2.14.4; Bürkner, 2018).

Bayesian mixed-effect models. For all Bayesian mixed-effect models, predictors were name agreement (high / low) and semantic context (homogeneous / heterogeneous), which were both contrast coded with (0.5, -0.5). The random effect structure for the models included random intercepts for participants and items, and did not include any random slopes because of the small number of observations (four per block) for each condition of each participant (Barr et al., 2013). Separate models were fitted for each dependent measure. All models had four chains and each chain had 4000 to 7000 iterations depending on model convergence (listed in model output tables). We used a warm-up (or burn-in) period of 1000 iterations in each chain, which means we removed the data based on the first 1000 iterations in order to correct the initial sampling bias.

All models used weak, widely spread priors that would be consistent with a range of null to moderate effects. The model of accuracy used family *bernoulli* combined with a

logit link, and the model used a student-*t* prior with 1 degree of freedom and a scale parameter of 2.5. The model of log-transformed utterance duration used a weak normal prior with an SD of 0.2, and the model of log-transformed total pause time had a weak normal prior with an SD of 1. Both were performed using the family *gaussian* combined with *identity* link. For chunk measures (i.e., total chunk number, first chunk length), the models had weak normal priors centered at zero with an SD of 3, and used the family *poisson* combined with the *log* link. In addition, the model of log-transformed onset latency used a weak normal prior with an SD of 0.2, and used the family *gaussian* combined with *identity* link. All models were run until the R hat value for each estimated parameter was 1.00, indicating full convergence. Analyses of posterior distributions given different prior distributions indicate that these priors were appropriate (see <https://osf.io/6jg4p/> for details).

For these models, the size of reported betas reflects estimated effect sizes, with larger absolute values of betas reflecting larger effects. We reported the parameters for which 95% Credible Intervals (hereafter, Cr.I) do not contain zero, which is analogous to the frequentist null hypothesis significance test: the parameter has a non-zero effect with high certainty. We also reported any parameters for which the point estimate for the beta is about twice the size of its error, as this also provides evidence for an effect: the estimated effect is large compared to the uncertainty around it. We also reported the posterior probability of the weak effects, indicating the proportion of samples with a value equal to or above the beta estimate.

Bayes factors. Bayes factors provide a way to quantify the evidence a data set provides in favor of one model over another. Although Bayes factors are defined on a continuous scale, several researchers have proposed to subdivide

Table 1. Means and standard deviations of the dependent variables calculated from trial onset (i.e., the start of the first picture) by name agreement and semantic context

	High name agreement		Low name agreement	
	homogeneous	heterogeneous	homogeneous	heterogeneous
Accuracy	85	88	87	91
Utterance duration (ms)	10424 (2628)	10152 (2560)	10960 (2636)	10762 (2621)
Total pause time (ms)	2579 (2012)	2339 (1991)	3022 (2049)	2855 (2007)
Total chunk number	5.3 (3.3)	5.1 (3.4)	6.1 (3.5)	5.8 (3.5)
First chunk length	5.2 (4.0)	5.2 (4.1)	4.3 (3.3)	4.5 (3.7)
Onset latency (ms)	1355 (385)	1312 (364)	1441 (447)	1415 (437)

Note. Standard deviations are given in parentheses. All time and chunking measures reflect fully correct trials only.

the scale in discrete evidential categories (e.g., M. D. Lee & Wagenmakers, 2014; Schönbrodt & Wagenmakers, 2018), which we report below. To obtain Bayes factors, we computed a series of reduced models eliminating each effect of interest one at a time, and then compared the reduced and full model using bridge sampling. These models used the same priors as the Bayesian mixed-effect models, but with a higher number of iterations, i.e., 20000. Sensitivity analyses suggest that the priors we selected were reasonable for this analysis, though they did have a moderate effect on the Bayes factor for the name agreement effect on log-transformed utterance duration (see <https://osf.io/6jg4p/> for details).

Analyses without the first row. Before data collection, we also planned to conduct an additional set of analyses where four dependent variables (i.e., accuracy, utterance duration, total pause time, and total chunk number) were calculated from the onset of naming the fifth picture (i.e., from the second row). This was done because the semantic context effect often arises from the second cycle (analogous to the second row of pictures in our study) and stays stable over subsequent cycles (Belke, 2017; Belke et al., 2005; Damian & Als, 2005).

Results

One participant was removed from further analyses because their responses were not recorded. The data from the remaining 40 participants was checked for errors, removing from analysis any trials with implausible names (e.g., *handschoen* 'glove' misnamed as *jas* 'coat'), hesitations (e.g., *sok* 'sock' named as *sss...sok*), self-corrections (e.g., *oor* 'ear' named as *neus...oor* 'nose...ear') and any trials where objects were omitted or named in the wrong order. Two more participants were then excluded because of high error rates (> 60%), following exclusion criteria we set before data collection. For the remaining 38 participants, the exclusion of inaccurate trials resulted in a loss of 12.17% of the data (range by participants: 0 - 37.5% of removed trials). Finally, any data points that were more than 2.5 standard deviations below or above the participant mean were removed for time measures (0.12 % for log-transformed utterance duration, 2.31% for log-transformed total pause time, and 0.81% for

log-transformed onset latency). Descriptive statistics of all dependent variables are shown in Table 1.

Accuracy. Participants produced the intended responses on 88% of the naming trials. As shown in Tables 1 and 2, a Bayesian mixed-effect model showed that accuracy was not influenced by name agreement, but it was considerably lower in the homogeneous condition than in the heterogeneous condition ($\beta = -0.379$, $SE = 0.188$, 95% Cr.I = [-0.753, -0.015]). Name agreement and semantic context did not interact. However, as shown in Table 3, Bayes factors showed only weak evidence in favor of the name agreement effect ($BF = 1.75$), and presented moderate evidence for the semantic context effect ($BF = 3.64$). There was only weak evidence against the interaction between name agreement and semantic context ($BF = 0.86$). In short, accuracy was somewhat affected by semantic context but not affected much by name agreement.

Utterance duration. As shown in Figure 2 and Table 2, a Bayesian mixed-effect model showed that log-transformed utterance duration was significantly longer for low name agreement pictures than for high name agreement pictures ($\beta = -0.055$, $SE = 0.018$, 95% Cr.I = [-0.091, -0.019]), but did not vary by semantic context. Name agreement and semantic context did not interact. Correspondingly, as shown in Table 3, Bayes factors showed moderate evidence in favor of the name agreement effect ($BF = 7.60$), but presented moderate evidence against the semantic context effect ($BF = 0.22$). There was moderate evidence against the interaction between name agreement and semantic context ($BF = 5.49$). In sum, utterance duration was affected by name agreement only.

Total pause time. As shown in Figure 2 and Table 2, a Bayesian mixed-effect model showed that log-transformed total pause time was longer for low name agreement pictures than for high name agreement pictures ($\beta = -0.254$, $SE = 0.057$, 95% Cr.I = [-0.366, -0.143]). There was moderate evidence for a semantic context effect ($\beta = 0.108$, $SE = 0.057$, 95% Cr.I = [-0.005, 0.22]). Note that while the 95% Cr.I contains zero, the point estimate is high relative to the error around it, and 97% of the posterior distribution around the estimated effect is above zero. This demonstrates that log-transformed total pause time was longer in the homogeneous than in the heterogeneous conditions.

Table 2. Results of Bayesian mixed-effect models for all dependent variables calculated from trial onset

		Estimate	Est.error	95% Cr. I		Effective samples
				lower	upper	
Accuracy						
	Intercept	2.27	0.177	1.936	2.636	3803
<i>Population-level effects</i>	Name Agreement	-0.309	0.186	-0.677	0.052	10504
	Semantic Context	-0.379	0.188	-0.753	-0.015	9697
	NA × SC	0.238	0.375	-0.5	0.972	9925
<i>Group-level effects</i>	Participant_sd(Intercept)	0.853	0.147	0.604	1.173	4228
	Item_sd(Intercept)	0.34	0.144	0.042	0.619	2278
Log-transformed utterance duration						
	Intercept	9.242	0.033	9.176	9.305	1593
<i>Population-level effects</i>	Name Agreement	-0.055	0.018	-0.091	-0.019	4057
	Semantic Context	0.024	0.018	-0.012	0.059	3865
	NA × SC	0.008	0.036	-0.063	0.078	3891
<i>Group-level effects</i>	Participant_sd(Intercept)	0.189	0.023	0.151	0.242	2526
	Item_sd(Intercept)	0.06	0.007	0.047	0.075	4494
Log-transformed total pause time						
	Intercept	7.633	0.101	7.435	7.839	552
<i>Population-level effects</i>	Name Agreement	-0.254	0.057	-0.366	-0.143	2703
	Semantic Context	0.108	0.057	-0.005	0.22	2581
	NA × SC	0.06	0.112	-0.162	0.282	2970
<i>Group-level effects</i>	Participant_sd(Intercept)	0.592	0.072	0.466	0.749	1382
	Item_sd(Intercept)	0.176	0.024	0.135	0.227	3224
Total chunk number						
	Intercept	1.62	0.075	1.475	1.769	654
<i>Population-level effects</i>	Name Agreement	-0.139	0.038	-0.214	-0.063	3889
	Semantic Context	0.045	0.038	-0.031	0.12	3597
	NA × SC	0.016	0.078	-0.135	0.174	3461
<i>Group-level effects</i>	Participant_sd(Intercept)	0.439	0.054	0.347	0.558	1331
	Item_sd(Intercept)	0.109	0.018	0.077	0.147	3944
First chunk length						
	Intercept	1.436	0.092	1.251	1.617	690
<i>Population-level effects</i>	Name Agreement	0.172	0.057	0.059	0.282	2749
	Semantic Context	-0.009	0.058	-0.122	0.102	2601
	NA × SC	0.052	0.115	-0.174	0.284	2730
<i>Group-level effects</i>	Participant_sd(Intercept)	0.533	0.067	0.418	0.682	1285
	Item_sd(Intercept)	0.182	0.024	0.14	0.234	3412
Log-transformed onset latency						
	Intercept	7.198	0.028	7.141	7.253	1130
<i>Population-level effects</i>	Name Agreement	-0.055	0.013	-0.079	-0.03	10977
	Semantic Context	0.025	0.013	-0.001	0.05	11029
	NA × SC	0.011	0.025	-0.038	0.06	11221
<i>Group-level effects</i>	Participant_sd(Intercept)	0.167	0.021	0.131	0.213	2336
	Item_sd(Intercept)	0.021	0.01	0.002	0.039	2835

Note. Models for log-transformed total pause time and total chunk number were run for 5000 iterations, model for log-transformed utterance duration was run for 7000 iterations, and models for other dependent variables were run for 4000 iterations. Bolded values indicate effects where the 95% Cr.I does not contain zero; Italicized values indicate effects where the beta estimate is twice the estimate of the standard error. NA refers to name agreement, SC refers to semantic context.

Table 3. Bayes factors for all dependent variables calculated from trial onset

	NA effect	SC effect	Null Interaction
Accuracy	1.75	3.64	0.86
Log-transformed utterance duration	7.60	0.22	5.49
Log-transformed total pause time	343.85	0.40	7.85
Total chunk number	6.34	0.03	38.32
First chunk length	1.55	0.02	24.34
Log-transformed onset latency	340.22	0.47	7.36

Note. NA refers to name agreement, SC refers to semantic context. Bolded values indicate moderate or above evidence in favor of the effects ($BF > 3$); Italicized values indicate moderate or above evidence against the effects ($BF < 1/3$); Regular values indicate only weak evidence in favor of or against the effects ($1/3 < BF < 3$).

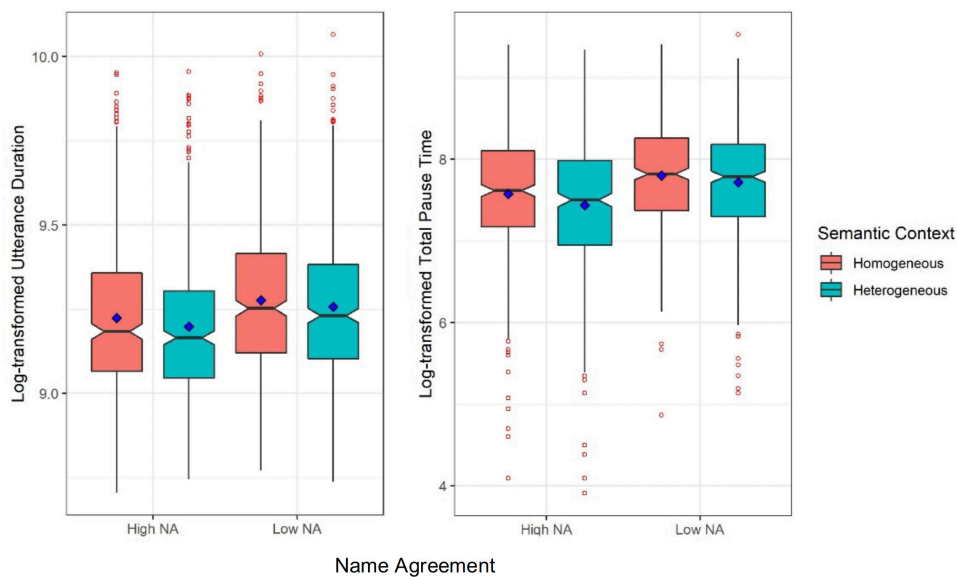


Figure 2. Log-transformed utterance duration (left) and log-transformed total pause time (right) calculated from trial onset split by name agreement (NA: high, low) and semantic context (homogeneous, heterogeneous)

Blue squares represent condition means and red points reflect outliers.

Again, name agreement and semantic context did not interact. Bayes factors showed a slightly different pattern: as shown in Table 3, Bayes factors showed extreme evidence in favor of the name agreement effect ($BF = 343.85$)³, but only weak evidence against the semantic context effect ($BF = 0.40$). There was moderate evidence against the interaction between name agreement and semantic context ($BF = 7.85$). Thus, consistent with the results of utterance duration, total pause time was affected by name agreement only.

Total chunk number. As shown in Figure 3 and Table 2, a Bayesian mixed-effect model showed that participants grouped their responses in more chunks for low name agreement pictures than high name agreement pictures (β

$= -0.139$, $SE = 0.038$, $95\% Cr.I = [-0.214, -0.063]$). Total chunk number was not impacted by semantic context, with no interaction between name agreement and semantic context. Bayes factors showed the same pattern, as shown in Table 3, with moderate evidence in favor of the name agreement effect ($BF = 6.34$), but moderate evidence against the semantic context effect ($BF = 0.03$). There was very strong evidence against the interaction between name agreement and semantic context ($BF = 38.32$). In sum, again, total chunk number was influenced by name agreement only.

First chunk length. As shown in Figure 3 and Table 2, a Bayesian mixed-effect model showed that participants planned fewer names in their first response chunk for low name agreement pictures than high name agreement pic-

³ Changing this prior to something less informative reduces this Bayes factor, but still shows strong or moderate evidence in favor of the effect. See <https://osf.io/6jg4p/> for details.

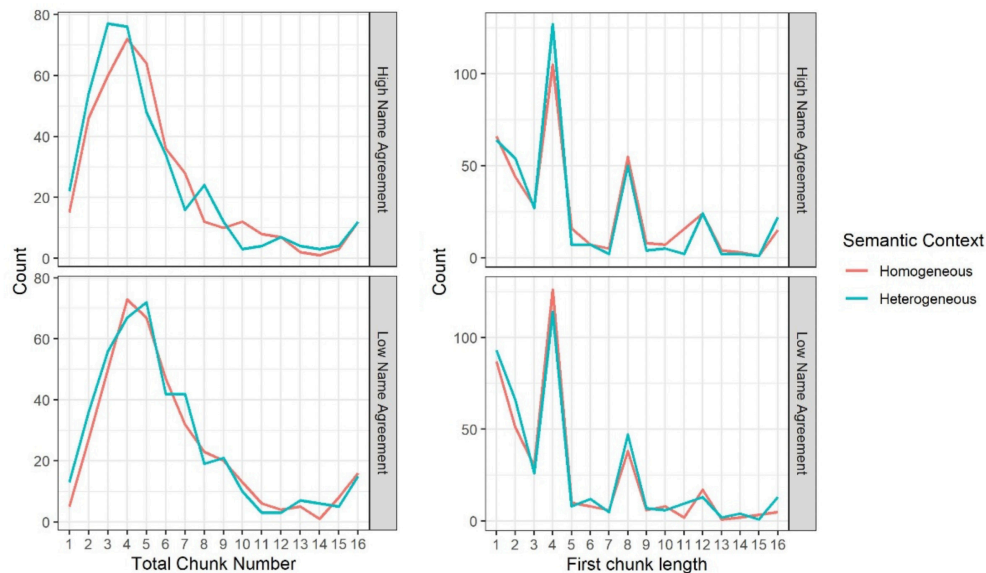


Figure 3. Total chunk number (left) and first chunk length (right) calculated from trial onset split by name agreement (high, low) and semantic context (homogeneous, heterogeneous)

tures ($\beta = 0.172$, $SE = 0.057$, 95% Cr.I = [0.059, 0.282]), but first chunk length was not impacted by semantic context and there was no interaction between name agreement and semantic context. As shown in Table 3, Bayes factors showed a matching pattern: only weak evidence in favor of the name agreement effect ($BF = 1.55$), and moderate evidence against the semantic context effect ($BF = 0.02$). There was strong evidence against the interaction between name agreement and semantic context ($BF = 24.34$). Thus, first chunk length appeared to depend on name agreement, but not semantic context.

Onset latency. As shown in Table 2, a Bayesian mixed-effect model showed that log-transformed onset latency was longer for low than high name agreement pictures ($\beta = -0.055$, $SE = 0.013$, 95% Cr.I = [-0.079, -0.03]). There was moderate evidence for a semantic context effect ($\beta = 0.025$, $SE = 0.013$, 95% Cr.I = [-0.001, 0.05]). Note that while the 95% Cr.I contains zero, the point estimate is high relative to the error around it, and 97% of the posterior distribution around the estimated effect is above zero. This demonstrates that log-transformed onset latency was longer in the homogeneous context than in the heterogeneous context. Name agreement and semantic context did not interact. Bayes factors showed a slightly different pattern: as shown in Table 3, Bayes factors showed extreme evidence in favor of the name agreement effect ($BF = 340.22$), but presented only weak evidence against the semantic context effect ($BF = 0.47$). There was moderate evidence against the interaction between name agreement and semantic context ($BF = 7.36$). Thus, the results observed for onset latency matched those obtained for the remaining dependent variables: name agreement had an impact, but semantic context did not.

Results from the onset of naming the fifth picture

Recall that earlier studies have showed that semantic context effects are typically not seen when the pictures of a set are named for the first time (e.g., Belke, 2017; Belke et al., 2005; Damian & Als, 2005). As shown in Figure 4, our results are, at least descriptively, consistent with this pattern. Semantic context effects were not present when participants named the first row of objects, but appeared in the following rows. Analyses for the data set without the first row were conducted to assess the semantic context effect from the second row onwards. As the results were largely comparable to the full data set, we only report differences from the main analyses. See Appendix B for full details of each analysis.

Bayesian mixed-effect models showed that semantic context did not influence accuracy, but affected log-transformed utterance duration ($\beta = 0.038$, $SE = 0.016$, 95% Cr.I = [0.006, 0.071]), log-transformed total pause time ($\beta = 0.17$, $SE = 0.075$, 95% Cr.I = [0.023, 0.318]), and total chunk number ($\beta = 0.070$, $SE = 0.034$, 95% Cr.I = [0.003, 0.136]) (see Table B1). However, Bayes factors slightly contradicted these analyses (see Table B2): There was only weak evidence in favor of semantic context effects on the time measures ($1/3 < BFs < 3$). There was moderate evidence against the semantic context effect on total chunk number ($BF = 0.10$). Thus, even when the first row was excluded from the analyses, there was at best weak evidence for semantic context effects on any of the dependent measures.

Post-hoc power analyses. To test whether the weak semantic context effects and null interaction were due to relatively small sample size in our study, we conducted a post-hoc power analyses at different sample sizes by using lme4 package (Bates et al., 2015) in R version 4.0.3 (R Core Team, 2020). For time measures, separate linear mixed-effect models with the same structure as the Bayesian mixed-effect

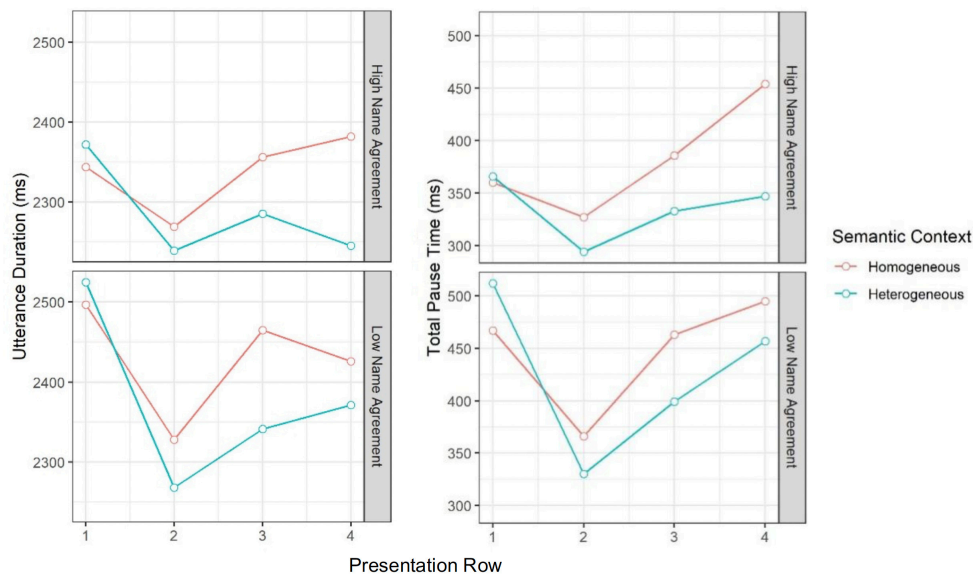


Figure 4. Utterance duration (left) and total pause time (right) in each row split by name agreement (high, low) and semantic context (homogeneous, heterogeneous)

fect models were performed. In each estimation, 86% of items (i.e., 40 trials) were included, and actual values of means and standard deviations in each condition were used. The number of simulations was 1000. To obtain power values, we compared the model with each effect of interest and the one without the effect (see <https://osf.io/6jg4p/> for details). As shown in Figure 5 (left), power values for the semantic context effects on time measures were relatively low for 38 participants (Powers < 0.5), while the values would be larger than 0.8 when testing at a minimum of 84 participants. This finding suggests that reliable semantic context effects can be detected for a large sample size. However, the power values for the interaction between name agreement and semantic context on time measures (see Figure 5, right) were extremely low even for a large enough sample size (e.g., Powers < 0.14 for 200 participants), which suggests that the null interaction cannot be attributed to the relatively small sample size in our study. Since the results for time measures calculated from the onset of naming the fifth picture are largely comparable to those from trial onset, we report them in Appendix C (see Figure C1).

Discussion

The present study investigated the feasibility of conducting spoken language production research in an online environment. We specifically explored the usefulness of measuring multiple dependent variables. We examined two previously studied effects related to lexical selection—semantic context and name agreement—in a modified blocked-cyclic naming paradigm. Six dependent variables were measured: naming accuracy, utterance duration, total pause time, total chunk number, first chunk length, and onset latency. We found strong evidence for name agreement effects, but little evidence for semantic context effects or interactions of the two variables. In this discussion, we comment on these findings, focusing primarily on their

methodological implications.

As predicted, we found robust name agreement effects on all measures except accuracy, with longer speech onset latencies, utterance durations and pause times, more response chunks, and shorter first chunk length for the naming of low name agreement pictures than high name agreement pictures. These results suggest that participants achieved lexical selection for the object names incrementally, at several time points during the process of multiple-object naming, and that they tended to plan their speech more sequentially with audible pauses between their responses when speech planning demand was high. These findings are important, as they suggest that measures of utterance durations and speech fluency can be exploited to study lexical access of speech production, in addition to, or instead of speech onset latencies. Of course, the sensitivity of utterance durations and speech fluency to the duration of cognitive processes underlying speech planning is not a new insight. For instance, some of the earliest theories of speech planning relied on analyses of pauses and disfluencies (e.g., Goldman-Eisler, 1972; Levelt, 1989), and, as described earlier, the RAN paradigm (Denckla & Rudel, 1976) that is often used in reading research measures total utterance durations (e.g., Araújo et al., 2020). The present study therefore may be seen as a reminder of the usefulness of these dependent variables to complement measurement of speech onset latencies. In interpreting experimental findings, it is, of course, always important to keep in mind that every dependent measure, be it speech onset latency or utterance duration, is likely to be affected by multiple influences. Speech onset latencies may, for instance, reflect not only on the time required to retrieve the first object name, but also on the time required for any advance planning of the following object names a participant may engage in. Similarly, total utterance durations will not only depend on the retrieval times for all object names but also on the strategies participants use to coordinate speech planning

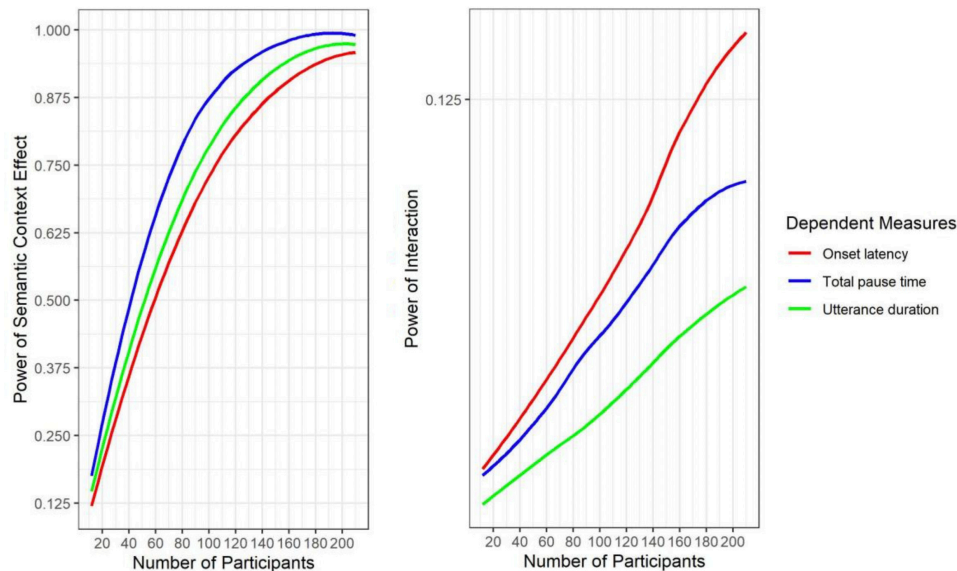


Figure 5. Results of post-hoc power analyses for the semantic context effects (left) and the interaction between name agreement and semantic context (right) on time measures calculated from trial onset

and speaking. Because speech planning can happen during articulation, utterance duration may be less sensitive to the effects of planning difficulty than onset latencies.

In this web-based paradigm we did, somewhat unexpectedly, observe robust evidence for name agreement effects on speech onset latencies, which replicates the effects of lab-based studies (e.g., Alario et al., 2004; Shao et al., 2014). Thus, our initial concern that speech onset latencies would be unreliable turned out to be unwarranted. Other recent studies using internet-based paradigms have provided similar evidence for the reliability of onset latencies, as they replicated several key findings of the speech production literature, including the word frequency effect (Fairs & Strijkers, 2021), the cumulative semantic interference effect (Stark et al., 2021), and the semantic interference effect in the picture-word interference paradigm (Vogt et al., 2021). Fairs and Strijkers (2021) compared the results of their web-based study to those of an otherwise identical study run in the laboratory. They found overall longer latencies in the web-based study but no difference in the size of the word frequency effect. Similarly, Stark and colleagues (2021) reported cumulative semantic interference effect comparable to effects found in earlier lab-based studies. In short, there is now good evidence that speech onset latencies can be recorded with good accuracy in web-based language production studies.

To return to our study, when the dependent variables were calculated from trial onset, semantic context only affected accuracy and total pause time. By contrast, when the dependent variables were calculated from the onset of naming the fifth picture (the first one in the second row), semantic context effects were found for all dependent variables except accuracy. This pattern is consistent with earlier lab-based studies using the classic blocked-cyclic naming paradigm (with one picture being displayed and named per trial) and showing that semantic context effects are only obtained from the second naming cycle onwards (e.g.,

Belke, 2017; Belke et al., 2005; Damian & Als, 2005). However, in our experiment, Bayes factors showed only weak evidence in favor of these semantic context effects on any measure except accuracy (BFs < 3). This suggests that the semantic context effects in our web-based study were relatively weak.

There are a number of reasons why the semantic context effects may have been weak. First, it could be that the simultaneous presentation of objects, compared to the sequential presentation, increased facilitatory conceptual or repetition priming effects and counteracted the inhibitory semantic context effects (as would be consistent with Abdel Rahman & Melinger, 2009; Howard et al., 2006; Oppenheim et al., 2010). This implies that semantic context effects might always be weak when the pictures are shown simultaneously. The effects of simultaneous versus successive presentation of pictures on the occurrence of semantic context effects should be further investigated. More generally, the timing of picture presentation (simultaneous, successive at a rapid or fast pace) may affect speakers' memory for the pictures already named and their planning for upcoming pictures, which should be kept in mind when designing a study.

Second, compared with onset latencies, measures of utterance durations and speech fluency during multiple object naming may be less sensitive to semantic context, or to any other variable affecting the speed of lexical access. Consistent with this proposal, Belke and Meyer (2007) found a robust semantic context effect on onset latencies, a small semantic context effect on word durations, but no effect on pause rates for the young speakers in their study. Semantic context effects may be hard to detect in measures of utterance durations and speech fluency because these measures depend not only on lexical access times, but also on multiple other variables, including the time required for phonetic planning, prosodic planning, and articulation, which may vary from trial to trial. Thus, while speech durations

and speech fluency can be exploited to assess the speed of word planning processes, subtle effects on word planning times may be obscured by other influences.

In addition, we found that semantic context did not interact with name agreement on any dependent variable, with Bayes factors showing moderate evidence or better (BFs > 3 for null interactions on all measures except accuracy). This might reflect that semantic context effects are not modulated by name agreement, suggesting that lexical selection can be achieved without competition, in line with the model proposed by Oppenheim and colleagues (2010). Alternatively, the interaction, just like the main effect of semantic context, may have been too subtle to be detected in analyses of utterance durations and speech fluencies.

A robust semantic context effect or an interaction between name agreement and semantic context may have been obtained with a larger sample size. We determined our sample size in terms of previous work: by collecting data from 41 participants, we doubled the number of participants tested in most lab-based semantic context experiments recording speech onset latencies (about 20 participants; e.g., Belke & Meyer, 2007; Damian & Als, 2005). A power simulation for determining sample size before the present study was not possible, as no comparable studies were available. However, we conducted post-hoc power calculations based on our results (see Figure 5), which suggest that robust semantic context effects indeed can be detected when testing at a minimum of 84 participants especially on total pause time. However, the interaction of semantic context and name agreement seems to be non-existent even for a large enough sample size (e.g., 200 participants). The results of post-hoc power analyses can now be used for a power simulation to estimate the sample size needed to observe effects of interest in future work.

In sum, we found strong evidence for name agreement effects, but weak evidence for semantic context effects. This pattern is consistent with the observation that name agreement effects on speech onset latencies tend to be descriptively larger than semantic context effects (e.g., Damian et al., 2001; Shao et al., 2014, 2015). Moreover, unlike semantic context effects, name agreement effects do not hinge on relationships between successive object names and consequently may be less likely to be affected by the timing of stimulus presentation.

Given the relative novelty of web-based studies of language production, we close by briefly commenting on the general quality of the data. It has been argued that the data quality of web-based experiments may be affected by poor compliance or distraction (e.g., Jun et al., 2017), and Fairs and Strijkers (2021) reported that 22% of their participants did not comply with the instructions. Other studies have shown no evidence for decreased attention and have demonstrated comparable data quality for web-based and lab-based studies (e.g., Casler et al., 2013; de Leeuw & Motz, 2016). Our results are consistent with the latter findings. There is little reason to assume that the participants in a web-based study will generally be less engaged or attentive than they would be in a laboratory setting. The speech recordings contained clearly articulated naming responses, no noise in the audio files, and little within-participant variation in the length of audio files per trial. Moreover,

we had a much lower rate of participant dropout than earlier web-based studies which reported dropout rates of over 30% (e.g., Sauter et al., 2020; Zhou & Fishbach, 2016). In our study, only 3 out of 41 participants (7.3%) were excluded from the analyses, one for technical reasons (the computer failed to record their speech responses) and two because they showed low overall accuracy (less than 40% correct responses). Unlike other web-based studies that used crowdsourcing marketplaces such as Amazon Mechanical Turk (e.g., Anwyl-Irvine et al., 2021; Schnoebelen & Kuperman, 2010), we recruited participants from the pool of individuals that we also use for lab-based studies. They are generally highly motivated and often have experience in participating in psycholinguistic studies. This most likely helped to ensure high-quality data collection. More generally, the success of an experiment, be it laboratory or web-based, depends on the adequate selection, instruction and motivation of the participants. There is no reason to assume that web-based experiments necessarily yield data of poorer quality than lab-based experiments do.

To conclude, the present study, along with several others, supports the feasibility of conducting spoken language production research on web-based platforms. Speech onset latencies turned out to be more reliable than we had assumed. Moreover, the durational properties of multi-word utterances such as utterance duration and speech fluency can be measured to examine processing times for lexical access. These measurements, therefore, are promising dependent variables for future spoken language production research with a modified blocked-cyclic naming paradigm, at least for research questions concerning variations in the speed and success of lexical access. Overall, this study supports the validity of the modified blocked-cyclic naming paradigm as one that is more similar to real-world speaking relative to a classic single picture naming paradigm.

Combined, the present study suggests that web-based studies are a promising addition or alternative to lab-based research. They can be used not only when there are travel restrictions or mobility issues for experimenters, but also to reach groups of participants who may be reluctant or unable to visit a lab. In short, they may contribute to rendering psycholinguistics a more inclusive field.

Contributions

Contributed to conception and design: Jieying He, Antje Meyer, Laurel Brehm.

Contributed to acquisition of data: Jieying He.

Contributed to analysis and interpretation of data: Jieying He, Ava Creemers, Laurel Brehm.

Drafted and/or revised the article: Jieying He, Antje Meyer, Ava Creemers, Laurel Brehm.

Approved the submitted version for publication: Jieying He, Antje Meyer, Ava Creemers, Laurel Brehm.

Acknowledgments

We would like to thank Maarten van den Heuvel for programming; Carlijn van Herpt for double checking stimuli;

Annelies van Wijngaarden, Inge Pasman, Dennis Joosen, Carlijn van Herpt, and Esther de Kerf for data coding.

Funding

This research was supported by the Max Planck Society. We also would like to express our gratitude to the China Scholarship Council for the support of the first author's study at the Max Planck Institute for Psycholinguistics.

Competing interests

The authors report no conflict of interest.

Supplemental material

Appendix A. Stimuli in the present study

Appendix B. Results for the analyses without the first row
Appendix C. Results of post-hoc power analyses on time measures calculated from the onset of naming the fifth picture

Data accessibility statement

All stimuli, participant data, and analysis scripts can be found on this paper's project page on the [OSF, <https://osf.io/6jg4p/>]

Submitted: May 05, 2021 PST, Accepted: November 01, 2021 PST



This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CCBY-4.0). View this license's legal deed at <http://creativecommons.org/licenses/by/4.0> and legal code at <http://creativecommons.org/licenses/by/4.0/legalcode> for more information.

REFERENCES

- Abdel Rahman, R., & Melinger, A. (2009). Semantic context effects in language production: A swinging lexical network proposal and a review. *Language and Cognitive Processes, 24*(5), 713–734. <https://doi.org/10.1080/01690960802597250>
- Alario, F.-X., Ferrand, L., Laganaro, M., New, B., Frauenfelder, U. H., & Segui, J. (2004). Predictors of picture naming speed. *Behavior Research Methods, Instruments, & Computers, 36*(1), 140–155. <https://doi.org/10.3758/bf03195559>
- Anwyl-Irvine, A., Dalmaijer, E. S., Hodges, N., & Evershed, J. K. (2021). Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behavior Research Methods, 53*(4), 1407–1425. <https://doi.org/10.3758/s13428-020-01501-5>
- Araújo, S., Huettig, F., & Meyer, A. S. (2020). What underlies the deficit in rapid automatized naming (RAN) in adults with dyslexia? Evidence from eye movements. *Scientific Studies of Reading, 25*(6), 534–549. <https://doi.org/10.1080/10888438.2020.1867863>
- 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*(3), 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Belke, E. (2017). The role of task-specific response strategies in blocked-cyclic naming. *Frontiers in Psychology, 7*, 1955. <https://doi.org/10.3389/fpsyg.2016.01955>
- Belke, E., & Meyer, A. S. (2007). Single and multiple object naming in healthy ageing. *Language and Cognitive Processes, 22*(8), 1178–1211. <https://doi.org/10.1080/01690960701461541>
- Belke, E., Meyer, A. S., & Damian, M. F. (2005). Refractory effects in picture naming as assessed in a semantic blocking paradigm. *The Quarterly Journal of Experimental Psychology Section A, 58*(4), 667–692. <https://doi.org/10.1080/02724980443000142>
- Boersma, P., & Weenink, D. (2009). *Praat: doing phonetics by computer (Version 5.1.05)*. <http://www.praat.org/>
- Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. W. (2020). The timing mega-study: Comparing a range of experiment generators, both lab-based and online. *PeerJ, 8*, e9414. <https://doi.org/10.7717/peerj.9414>
- Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal, 10*(1), 395–411. <https://doi.org/10.32614/rj-2018-017>
- Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior, 29*(6), 2156–2160. <https://doi.org/10.1016/j.chb.2013.05.009>
- Cooke, M., Barker, J., Lecumberri, M. L. G., & Wasilewski, K. (2011). Crowdsourcing for word recognition in noise. In P. Cosi, R. De Mori, G. Di Fabbrizio, & R. Pieraccini (Eds.), *Proceedings of Interspeech 2011* (pp. 3049–3052). International Speech Communication Association.
- Damian, M. F., & Als, L. C. (2005). Long-lasting semantic context effects in the spoken production of object names. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(6), 1372–1384. <https://doi.org/10.1037/0278-7393.31.6.1372>
- Damian, M. F., Vigliocco, G., & Levelt, W. J. M. (2001). Effects of semantic context in the naming of pictures and words. *Cognition, 81*(3), B77–B86. [https://doi.org/10.1016/s0010-0277\(01\)00135-4](https://doi.org/10.1016/s0010-0277(01)00135-4)
- de Leeuw, J. R., & Motz, B. A. (2016). Psychophysics in a Web browser? Comparing response times collected with JavaScript and Psychophysics Toolbox in a visual search task. *Behavior Research Methods, 48*(1), 1–12. <https://doi.org/10.3758/s13428-015-0567-2>
- Denckla, M. B., & Rudel, R. G. (1976). Rapid 'automatized' naming (R.A.N.): Dyslexia differentiated from other learning disabilities. *Neuropsychologia, 14*(4), 471–479. [https://doi.org/10.1016/0028-3932\(76\)90075-0](https://doi.org/10.1016/0028-3932(76)90075-0)
- Duñabeitia, J. A., Crepaldi, D., Meyer, A. S., New, B., Pliatsikas, C., Smolka, E., & Brysbaert, M. (2018). MultiPic: A standardized set of 750 drawings with norms for six European languages. *Quarterly Journal of Experimental Psychology, 71*(4), 808–816. <https://doi.org/10.1080/17470218.2017.1310261>
- Fairs, A., & Strijkers, K. (2021). Can we use the internet to study speech production? Yes we can! Evidence contrasting online versus laboratory naming latencies and errors. *PsyArXiv*. <https://doi.org/10.31234/osf.io/2bu4c>
- Ferreira, F., & Swets, B. (2017). The production and comprehension of resumptive pronouns in relative clause "island" contexts. In A. Cutler (Ed.), *Twenty-First Century Psycholinguistics: Four Cornerstones* (pp. 263–278). Routledge.
- Goldman-Eisler, F. (1972). Pauses, clauses, sentences. *Language and Speech, 15*(2), 103–113. <https://doi.org/10.1177/002383097201500201>
- Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science, 11*(4), 274–279. <https://doi.org/10.1111/1467-9280.00255>
- Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics, 38*(4), 555–568. <https://doi.org/10.1016/j.wocn.2010.08.002>

- Howard, D., Nickels, L., Coltheart, M., & Cole-Virtue, J. (2006). Cumulative semantic inhibition in picture naming: experimental and computational studies. *Cognition*, 100(3), 464–482. <https://doi.org/10.1016/j.cognition.2005.02.006>
- Jun, E., Hsieh, G., & Reinecke, K. (2017). Types of motivation affect study selection, attention, and dropouts in online experiments. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW), 1–15. <https://doi.org/10.1145/3134691>
- Kandel, M., Wyatt, C., & Phillips, C. (2021, March 4). *Transitioning to online language production: a direct comparison of in-lab and web-based experiments*. The 34th Annual CUNY Conference on Human Sentence Processing, Philadelphia, United States. https://www.cuny2021.io/wp-content/uploads/2021/02/CUNY_2021_abstract_131.pdf
- Konopka, A. E. (2019). Encoding actions and verbs: Tracking the time-course of relational encoding during message and sentence formulation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 45(8), 1486–1510. <https://doi.org/10.1037/xlm0000650>
- Lee, E.-K., Brown-Schmidt, S., & Watson, D. G. (2013). Ways of looking ahead: Hierarchical planning in language production. *Cognition*, 129(3), 544–562. [http://doi.org/10.1016/j.cognition.2013.08.007](https://doi.org/10.1016/j.cognition.2013.08.007)
- Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.
- Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. The MIT Press.
- Meyer, A. S., Wheeldon, L., van der Meulen, F., & Konopka, A. (2012). Effects of speech rate and practice on the allocation of visual attention in multiple object naming. *Frontiers in Psychology*, 3, 39. <https://doi.org/10.3389/fpsyg.2012.00039>
- Momma, S., & Ferreira, V. S. (2019). Beyond linear order: The role of argument structure in speaking. *Cognitive Psychology*, 114, 101228. <https://doi.org/10.1016/j.cogpsych.2019.101228>
- Mortensen, L., Meyer, A. S., & Humphreys, G. W. (2008). Speech planning during multiple-object naming: Effects of ageing. *Quarterly Journal of Experimental Psychology*, 61(8), 1217–1238. <https://doi.org/10.1080/17470210701467912>
- Nicenboim, B., & Vasishth, S. (2016). Statistical methods for linguistic research: Foundational ideas—Part II. *Language and Linguistics Compass*, 10(11), 591–613. <https://doi.org/10.1111/lnc3.12207>
- Oppenheim, G. M., Dell, G. S., & Schwartz, M. F. (2010). The dark side of incremental learning: A model of cumulative semantic interference during lexical access in speech production. *Cognition*, 114(2), 227–252. <https://doi.org/10.1016/j.cognition.2009.09.007>
- R Core Team. (2020). *R: A language and environment for statistical computing (Version 4.0.3)*. <http://www.R-project.org>
- Reimers, S., & Stewart, N. (2015). Presentation and response timing accuracy in Adobe Flash and HTML5/JavaScript Web experiments. *Behavior Research Methods*, 47(2), 309–327. <https://doi.org/10.3758/s13428-014-0471-1>
- Sauter, M., Draschkow, D., & Mack, W. (2020). Building, hosting and recruiting: A brief introduction to running behavioral experiments online. *Brain Sciences*, 10(4), 251. <https://doi.org/10.3390/brainsci10040251>
- Schnoebelen, T., & Kuperman, V. (2010). Using Amazon Mechanical Turk for linguistic research. *Psihologija*, 43(4), 441–464. <https://doi.org/10.2298/psi1004441s>
- Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, 25(1), 128–142. <https://doi.org/10.3758/s13423-017-1230-y>
- Shao, Z., Roelofs, A., Acheson, D. J., & Meyer, A. S. (2014). Electrophysiological evidence that inhibition supports lexical selection in picture naming. *Brain Research*, 1586, 130–142. <https://doi.org/10.1016/j.brainres.2014.07.009>
- Shao, Z., Roelofs, A., Martin, R. C., & Meyer, A. S. (2015). Selective inhibition and naming performance in semantic blocking, picture-word interference, and color-word Stroop tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41(6), 1806–1820. <https://doi.org/10.1037/a0039363>
- Stark, K., van Scherpenberg, C., Obrig, H., & Rahman, R. A. (2021). Web-based language production experiments: Semantic interference assessment is robust for spoken and typed response modalities. *PsyArXiv*. <https://doi.org/10.31234/osf.io/5k8de>
- van Casteren, M., & Davis, M. H. (2006). Mix, a program for pseudorandomization. *Behavior Research Methods*, 38(4), 584–589. <https://doi.org/10.3758/bf03193889>
- van Paridon, J., & Thompson, B. (2021). subs2vec: Word embeddings from subtitles in 55 languages. *Behavior Research Methods*, 53(2), 629–655. <https://doi.org/10.3758/s13428-020-01406-3>
- Vitkovitch, M., & Tyrrell, L. (1995). Sources of disagreement in object naming. *The Quarterly Journal of Experimental Psychology Section A*, 48(4), 822–848. <https://doi.org/10.1080/14640749508401419>
- Vogt, A., Hauber, R. C., Kuhlén, A. K., & Abdel Rahman, R. (2021). Internet based language production research with overt articulation: Proof of concept, challenges, and practical advice. *PsyArXiv*. <https://doi.org/10.31234/osf.io/cyvwf>
- Withers, P. (2017). *Frinex: Framework for interactive experiments*. Zenodo. <https://doi.org/10.5281/ZENODO.3522911>
- Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, 111(4), 493–504. <https://doi.org/10.1037/pspa0000056>

SUPPLEMENTARY MATERIALS

Peer Review History

Download: https://collabra.scholasticahq.com/article/29935-conducting-language-production-research-online-a-web-based-study-of-semantic-context-and-name-agreement-effects-in-multi-word-production/attachment/75732.docx?auth_token=aKP_-FdZ7ZpaKSFmGt7U

Supplemental Material Appendices

Download: https://collabra.scholasticahq.com/article/29935-conducting-language-production-research-online-a-web-based-study-of-semantic-context-and-name-agreement-effects-in-multi-word-production/attachment/75733.docx?auth_token=aKP_-FdZ7ZpaKSFmGt7U
