Typing time PWI: A scalable paradigm for studying lexical production

Shiva Upadhye and Richard Futrell
Department of Language Science
University of California, Irvine
{upadhyes, rfutrell}@uci.edu

Abstract

Lexical production research has relied extensively on in-person naming experiments such as Picture Word Interference (PWI). While the interference effects observed in the standard paradigm have been well-attested, PWI experiments have traditionally been underpowered, owing to limitations associated with large-scale data collection. In this study, we validate a scalable, typing time-based PWI paradigm that can be deployed on the internet, and enables large-scale replications of the interference effects observed in the spoken modality. We also propose an automated response coding process that incorporates production errors such as incorrect responses and restarts, and can be leveraged to illuminate aspects of response conflict and correction processes in incremental production.

Keywords: lexical production; Picture Word Interference, typing modality

Introduction

A key phenomenon for the study of language production at the level of individual words is lexical interference, in which a speaker ought to produce one word but mistakenly produces another instead. The widely-used Picture Word Interference (PWI) paradigm (Rosinski, 1977) probes this effect in a picture-naming setting, by presenting participants with a stimulus composed of a target image superimposed with an orthographic distractor, as shown in Figure 1. Various interference effects can be observed in this paradigm: for example, there is a general interference effect consisting of a delay in naming when the target is coupled with an incongruent distractor (Rosinski et al., 1975; W. R. Glaser & Dünghoff, 1984; MacLeod, 1991), and a semantic interference effect, characterized by a larger delay when the distractor is semantically related to the word than when it is unrelated (Lupker, 1979; Roelofs & Piai, 2017). Framing lexical production as a two-alternative forced choiced (2AFC) task, the PWI paradigm engages meta-cognitive processes such as executive control and conflict monitoring, similar to the Stroop task (Van Maanen & Van Rijn, 2010; Xiao et al., 2010; Starreveld & La Heij, 2017). Consequently, the paradigm is a workhorse method for studying how language processing mechanisms interact with cognitive control, for example as in the case of code-switching performance in bilingualism (Ehri & Ryan, 1980).

While the PWI paradigm is widely recruited in psycholinguistic research, its reach is hindered by the practicalities of running PWI experiments: participants must be recruited to participate in an in-person naming experiment; responses are recorded; the recordings must be coded and filtered; etc. As a result, PWI experiments are often limited in terms of the number of subjects and the diversity of participant population. The lack of scalability contributes to weak statistical power for detecting moderators of interference (Bürki et al., 2020). Whereas recent web-based PWI experiments have addressed some of the concerns associated with running participants in-person (Fairs & Strijkers, 2021; Vogt et al., 2022), transforming the audio recordings into usable reaction time (RT) measures still requires a substantial data processing effort, which also includes judgement calls such as assessing the quality of the recording and verifying the validity of the participants’ responses. Consequently, automating data processing and scaling up the experiment is costly, both in terms of experimenter effort and resources.

We present a scalable, internet-based version of the PWI paradigm based on typing time, which we call tPWI. The paradigm not only supports fully automated extraction of a variety of dependent measures (not just onset latency), it also enables automatic characterization of the different kinds of errors that may be encountered in the participants’ responses. We demonstrate that the tPWI paradigm not only replicates the general and semantic interference effects observed in the spoken paradigm, but also those associated with well-attested variations such as manipulations of stimulus onset asynchrony (SOA) and response set membership. Furthermore, we also show that incorporating different types of responses including production errors facilitates unified analyses of response times and error rates in the tPWI paradigm.

Related Work

PWI paradigm and manipulations

A critical question common to both Stroop task and PWI research is whether the interference arises at the level of perceptual processing, lexical selection, or articulatory planning (M. Glaser & Glaser, 1982). Two accounts of the lexical interference effect have gained traction in PWI literature (Bürki et al., 2020). The first account, known as the competitive lexical selection hypothesis, argues that semantic interference...
arises because related distractors accrue a higher degree of activation than their unrelated counterparts by virtue of shared semantic features, hence making them likely candidates for production (Levelt et al., 1999; Damian & Bowers, 2003; Abdel Rahman & Melinger, 2009). The response exclusion account, on the other hand, argues that the target and distractor do not necessarily compete for selection during lexical access. Rather, the high degree of activation received by a semantically related distractor activates its phonological form, which then enters the articulatory buffer as a candidate for production (Mahon et al., 2007). Hence, under this account, interference originates at the level of articulation rather than retrieval.

Several experimental manipulations have been proposed in order to tease apart the predictions of the two accounts. Since the PWI task involves cross-modal stimuli, specifically a pictorial target and an orthographic distractor, the degree of semantic interference may be further modulated by temporal differences in target and distractor processing (M. Glaser & Glaser, 1982). In order to test for these effects, the timing of distractor presentation may be varied such that it is presented before (negative SOA), after (positive SOA), or at the same time (SOA = 0) as the target (W. R. Glaser & Düngelhoff, 1984). The Stimulus Onset Asynchrony (SOA) manipulation has been well-attested in the PWI literature. Particularly, the semantic interference effect has been observed at shorter SOAs since the target and distractor are more likely to be processed simultaneously (W. R. Glaser & Düngelhoff, 1984; Starreveld & Heij, 1996). In contrast, long negative SOAs have been associated with a speedup in naming or facilitation (Bloem et al., 2004; Zhang et al., 2016). While the distractor is activated much earlier than the target, the activation received by the distractor decays due to the long SOA interval. Consequently, some accounts ascribe facilitation to residual activation from the distractor (Finkbeiner & Carmazza, 2006) while others attribute it to internal monitoring and control processes that take scope after lexical selection (Python et al., 2018). Intriguingly, interactions between SOA and distractor condition have been admitted as evidence for both the competitive lexical selection and response exclusion accounts (Bürki et al., 2020).

Another critical manipulation that has been examined in PWI research is whether the words that appear as orthographic distractors also appear as targets in other trials. Since prior retrieval of a word typically primes the word for subsequent access, including distractors in the response set may affect their accessibility during production. This modulating effect of response set membership has been rigorously debated in the PWI literature: whereas some experiments report diminished interference or even facilitation for distractors not in the response set (Roelofs, 1992a), others report no significant difference in the magnitude of interference (Caramazza & Costa, 2001).

Production Experiments in the typing modality

The widespread use of typing across a diversity of communicative contexts has transformed it into a production modality worthy of further investigation. Whereas writing has traditionally been neglected in language production research due to its limited scope, function, and register, typing hews closer to speech since it lends itself to rapid execution and correction. Owing to its usage in naturalistic and conversational contexts, typing is subject to many of the same constraints that shape real-time speech production. Language production experiments involving typing, including in picture-naming paradigms, have yielded results that mirror patterns in oral speech production, including showing effects of frequency, contextual predictability, concreteness, and cumulative semantic interference (Cohen Priva, 2010; Torrance et al., 2018; Chen et al., 2021). In spite of the differences in the motor architectures engaged in both modalities, Pinet & Nozari (2018) found that ‘finger twister’ sequences, i.e. the typewritten analog of tongue twisters, produced errors that closely resembled those found in speech. Similarly, monitoring and error detection processes in both modalities have been associated with common electrophysiological signatures (Pinet & Nozari, 2020), although a proclivity toward correction rather than error prevention has been observed in typed production (Crump & Logan, 2013). More closely related to this work, Stark et al. (2021) find that the cumulative interference effect observed in a web-based blocked cyclic naming task is preserved in typed production.

Furthermore, in existing successful computational models, the representations and processes recruited by spoken and typed production systems share crucial features. Particularly, models of speech and typing are both characterized by two stages: the first involves the activation of lexico-semantic representations that, in turn, activate the constituent phonetic or graphemic schemata (Levelt, 1992; McClelland & Rumelhart, 1981), and the second involves a mapping between the activated schemata and modality-specific motor processes (Dell & O’Seaghdha, 1992; Crump & Logan, 2010).

Based on these findings, we take the general stance that typing data reflects many of the same processes as oral speech, plus some additional processes primarily involving motor control of the digits and factors such as keyboard layouts. As a result, results involving typing time can be indicative of underlying language production processes.

Web-based experimental setup

We design and host web-based tPWI using the Javascript-based Ibex platform (Zehr & Schwarz, 2018) which records all input keystrokes and their timing. Participant recruitment is handled through Prolific, and two screening criteria were applied: participants had to be native English speakers residing in the United States and could only use devices with physical keyboards to complete this experiment. To address issues that could arise due to the absence of experimenter interaction or supervision during the experiment, we introduce a rigorous
two-stage familiarization phase before the main experiment begins. In the first stage, participants view all the images in the response set paired with their canonical names. In the next stage, participants are asked to recall and type the name of all the images in a randomized order. Following practice trials where the participants are asked to name the image, they proceed to the experimental trials, where they are presented with the stimulus 300 ms after a fixation cross. The time at which the stimulus was rendered as well as the time corresponding to each keystroke was logged in order to compute response latency (see Fig 2 for illustration of the experiment design).

**Response validity and automated coding**

During the experiment, participants’ responses in an individual trial are evaluated online after they click the Enter button on their keyboard or press Enter button displayed on the screen. A response is considered valid if the first and last segments of the typed response match either the target or the distractor. In this case, the participant is allowed to proceed to the next trial. While the participants were instructed to type the canonical name of the image, distractors were deemed valid responses for two reasons. First, from a logistical standpoint, this allowed the participants to move to the next trial quickly, similar to spoken experiment. More importantly, the rate of distractor responses offers crucial insights into the implicit speed-accuracy trade-off at the participant level as well as the modality level. However, if the response is deemed invalid (sample response provided in Table 1), the participant is notified of their error and asked to correct their response in order to move to the next trial. To circumvent instances where a participant stays on a trial for too long because they failed to provide any valid response, a “Give up” button appears after the first attempt, allowing the participant to move to the next trial.

After data-collection, responses were categorized as fluent, edited, restarted, or invalid during post-processing\(^1\) (see Fig. 3). Responses that passed the online check in the first attempt and involved no edits (as measured by the backspace key) were automatically coded as fluent. A further distinction was made between fluent-correct responses and fluent-incorrect responses based on whether the participant typed the target or distractor (see Table 1). In contrast, responses where the participants attempt to correct minor ty-

\(^1\)Preprocessed data from all three experiments can be accessed at https://osf.io/z79u5/

<table>
<thead>
<tr>
<th>Category</th>
<th>Typed response</th>
<th>Target</th>
<th>Distractor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluent</td>
<td>c,h,u,r,c,h,(E)</td>
<td>church</td>
<td>castle</td>
</tr>
<tr>
<td></td>
<td>p,l,a,t,e,(E)</td>
<td>glass</td>
<td>plate</td>
</tr>
<tr>
<td>Edited</td>
<td>w,a,r,f,(B),d,z,o,s,(B),h,e,(E)</td>
<td>wardrobe</td>
<td>wardrobe</td>
</tr>
<tr>
<td></td>
<td>p,i,s,t,o,l,e,(E),(E),(B),(E)</td>
<td>pistol</td>
<td>bed</td>
</tr>
<tr>
<td>Restarted</td>
<td>b,(B),d,z,s,k,(E)</td>
<td>desk</td>
<td>bed</td>
</tr>
<tr>
<td></td>
<td>c,o,(B),(B),j,u,g,(E)</td>
<td>jug</td>
<td>cup</td>
</tr>
<tr>
<td></td>
<td>g,u,(B),(B),(B),(B),(B),(E)</td>
<td>pistol</td>
<td>cannon</td>
</tr>
</tbody>
</table>

Table 1: Sample typed responses from each response category. (B) means backspace, (E) means Enter.
pographic errors before completing their attempt were categorized as edited. Responses that initially failed the online check due to due to typographic errors were also included in this category, provided the response was within an edit distance of 2 before the user clicked the Enter key. We define invalid responses as those where the response failed the initial online check after the participant clicked the Enter key, and subsequent attempts aimed at revising the response rather than merely correcting typos. To distinguish between the two strategies, we match both the original and revised input to the two possible responses (i.e., the target and the distractor). If the original and corrected input matched different responses, the response was coded as invalid. If not, it was categorized as edited. Finally, we consider invalid responses to be different than restarted responses. We identify restarted responses as those where the participant’s initial typed input matched one response but was corrected to a different response before they completed an attempt i.e., clicked Enter. We accommodate this distinction in our taxonomy since a correction ensues before the user receives any overt feedback about the error, unlike in the case of invalid responses where the participants revise their response after being notified of an error.

Experiments

We replicate Roelofs & Piai (2017), a high powered spoken experiment, in the typing paradigm. Experiment 1 serves as a baseline replication, where all distractors are presented synchronously and belong to the response set. A critical difference between the typed replication and Roelofs & Piai (2017) is that the latter included three exposures for each of their items, whilst we include a single exposure. In Experiment 2, we manipulate the stimulus onset asynchrony, presenting distractors at three SOAs: -150 ms, 0 ms, and 150 ms. Finally, in Experiment 3, we present distractors synchronously but manipulate the response set membership by including incongruent (related and unrelated) distractors that do not appear as targets.

Proposed analyses We operationalize onset latency as the delay between when the target image was rendered and the first keystroke corresponding to either the target or distractor was registered. To hew close to the standard paradigm, we only consider the fluent trials where the participants typed the target. We test for two attested interference effects: a general interference (GI) effect, defined by longer latencies for incongruent trials relative to neutral trials, and a semantic interference (SI) effect, characterized by longer latencies for related distractors compared to unrelated distractors. We use linear mixed effects models (LMEMs) (Baayen et al., 2008) with random slopes and intercepts for subjects, targets, and distractors. For all models, the experimental condition was treated as a critical fixed effect. To test for the GI effect, the congruent condition was coded as 0 and the incongruent condition was coded as +1. Similarly in order to test for the SI effect, the unrelated condition was coded as the reference level while the related condition was coded as +1. For experiments 2 and 3, we include a condition × SOA term and a condition × distractor membership term respectively to test for the interaction between these two effects.

Experiment 1: Basic Semantic Interference

Materials and Participants: We use the English glosses of the Dutch stimuli from Roelofs & Piai (2017) to validate the paradigm. We invert the white-on-black line drawings from the original experiment into black-on-white images that are better suited for a web-based experiment set against a white background. As in the original experiment, each target image was paired with a distractor word that was (i) identical, (ii) semantically related, (iii) unrelated, and (iv) neutral, as indicated by an ‘XXXXX’ string. For this experiment, 100 native English speakers using devices with physical keyboards were recruited.

Results: In order to hew close to the spoken paradigm, we restrict the naming latency analyses to trials with no typos or corrections. A linear mixed effects model was used to model onset latency with condition as the fixed effect, and subjects and targets as random effects (Barr et al., 2013). Our analysis confirms a significant general interference effect, with slower typing onset latencies for the incongruent conditions relative to the neutral condition ($\beta = 154.26, SE = 17.14, p < 0.001$). Considering the subset of incongruent trials, we also find that the paradigm replicates the semantic interference effect ($\beta = 68.8302, SE = 26.31, p < 0.01$), as characterized by significantly slower latencies in the related versus unrelated condition (see Fig. 4a).

Experiment 2: Stimulus Onset Asynchrony

Materials and Participants: We use images from the validation experiment as stimuli but vary the timing of the distractor presentation with respect to the target. In particular, we present the distractor (i) synchronously (0 ms), (ii) shortly before the target (−150 ms), and (iii) shortly after the target (+150 ms). To ensure precision transitions between image frames, we convert the stimuli into the graphical interchange (GIF) format. The trial structure was identical to that in the previous experiment, and 96 participants were recruited using Prolific.

Results: We find a significant general interference effect for both short negative ($\beta = 103.30, SE = 23.60, p < 0.001$) and short positive SOAs ($\beta = 105.5114, SE = 25.3127, p < 0.001$), with a significant delays in naming latency in the incongruent conditions (see Fig. 4b). We also recover a significant effect of semantic relatedness in the short positive SOA of 150 ms ($\beta = 58.5926, SE = 21.59, p < 0.01$). However, we do not find a significant delay in the related condition relative to the unrelated condition in the short negative SOA trials ($\beta = 29.93, SE = 16.1411, p > 0.1$). Examining the effects of SOA, we find that onset latencies are modulated by SOA, but only when distractors are presented -150 ms earlier than the target, as evinced by longer onset latency.
latencies in the negative SOA condition compared to the synchronous condition (Figure 4(b)). Compared to trials with SOA = 0, trials with SOA = -150 were characterized by a smaller difference between related and unrelated conditions ($\beta = -42.79, SE = 16.7927, p < 0.05$). Whereas a similar trend was observed for trials where distractors were presented 150 ms after the target, the difference was not found to be significant ($\beta = -16.09, SE = 31.5020, p > 0.5$).

**Experiment 3: Response set membership**

**Materials and Participants:** We use the same stimuli as Experiment 1, but include a response set manipulation by adding related and unrelated distractor words that do not appear as targets. Frequencies of these distractors were estimated using the SUBTLEXus corpus (Brysbaert & New, 2009) to control for distractor frequency effects. For this experiment, we recruited 60 native English speakers from Prolic. Four participants were excluded from analysis due to data-collection issues.

**Results:** Incongruent distractors not in the response set elicited a delay in naming relative to neutral condition, as evinced by a significant general interference effect ($\beta = 83.01, SE = 7.45, p < 0.001$). Related distractors that were not part of the response set were also characterized by longer naming delays ($\beta = 21.35, SE = 7.4962, p < 0.01$). For the unrelated trials, onset latencies were longer when the distractor was in the response set, although the difference was not found to be significant ($\beta = -8.49, SE = 7.85, p > 0.1$). However, in case of related distractors, response set membership had a significant effect on onset latency, with distractors in the response set eliciting greater delays in naming ($\beta = -31.32, SE = 12.4227, p < 0.05$).

**Toward integrated analyses of reaction times and errors**

We use the automated response coding process detailed in Figure 3 to identify and extract restarted responses (e.g., c,u,(Backspace),(Backspace), j,u,g) as well as incorrect responses i.e., trials where the participant typed the distractor rather than the target. For restarted responses, we omit control keys such as Backspace from our calculation, and compute response onset latency by identifying the timing of the first keystroke matching the restarted response. Word initial typos or characters not matching the distractor were considered *edited* rather than *restarted* responses. The per-category response rates for the tPWI validation experiment are summarized in Table 2.

<table>
<thead>
<tr>
<th>response</th>
<th>% of trials</th>
</tr>
</thead>
<tbody>
<tr>
<td>fluent</td>
<td>88.32%</td>
</tr>
<tr>
<td>edited</td>
<td>8.57%</td>
</tr>
<tr>
<td>restarted</td>
<td>0.64%</td>
</tr>
<tr>
<td>incorrect</td>
<td>0.86%</td>
</tr>
<tr>
<td>invalid</td>
<td>1.65%</td>
</tr>
</tbody>
</table>

Table 2: Percentage of total responses by category

Figure 5 illustrates the differences in onset latencies across response categories. To estimate the effects of response category, we fit a LMEM with onset latency as the dependent measure, response category as a critical fixed effect, and targets, distractors, and subjects as random effects. We observe that onset latencies of incorrect trials were faster than those of correct trials ($\beta = -227.74, SE = 61.18, p < 0.001$). Unlike correct trials, however, no significant difference between related and unrelated trials was found within incorrect trials ($\beta = 93.62, SE = 85.07, p > 0.1$). Intriguingly, a qualitative difference between the related and unrelated condition was observed for restarted trials, with slower restart latencies for unrelated rather than related trials ($\beta = -465.5891, SE = 145.1418, p < 0.01$).

**Discussion**

The goal of this study is not to adjudicate between different accounts of the lexical interference effects proposed in the
The authors are grateful to Ardi Roelofs and Vitoria Piai for kindly sharing the materials from Roelofs & Piai (2017).
References


