Behavior Research Methods (2018) 50:2173–2183
https://doi.org/10.3758/s13428-018-1040-9

The (un)reliability of item-level semantic priming effects

Tom Heyman[1] · Anke Bruninx[1] · Keith A. Hutchison[2] · Gert Storms[1]

Published online: 5 April 2018

Abstract

Many researchers have tried to predict semantic priming effects using a myriad of variables (e.g., prime–target associative strength or co-occurrence frequency). The idea is that relatedness varies across prime–target pairs, which should be reflected in the size of the priming effect (e.g., *cat* should prime *dog* more than *animal* does). However, it is only insightful to predict item-level priming effects if they can be measured reliably. Thus, in the present study we examined the split-half and test–retest reliabilities of item-level priming effects under conditions that should discourage the use of strategies. The resulting priming effects proved extremely unreliable, and reanalyses of three published priming datasets revealed similar cases of low reliability. These results imply that previous attempts to predict semantic priming were unlikely to be successful. However, one study with an unusually large sample size yielded more favorable reliability estimates, suggesting that big data, in terms of items and participants, should be the future for semantic priming research.

Keywords Semantic priming · Split-half reliability · Test–retest reliability · Semantic memory

Since Meyer and Schvaneveldt's (1971) seminal article, semantic priming has been the topic of much research in (cognitive) psychology and linguistics.[1] One of the main reasons for its popularity is the belief that priming offers unique insight into how people's semantic memory is organized. That is, semantic priming effects—the improvements in speed and accuracy when responding to a target stimulus (e.g., *dog*) that is preceded by a related prime stimulus (e.g., *cat*)—supposedly reflect the structure of the mental lexicon (see McNamara, 2005, for a review). As such, many (recent) studies have linked priming to various prime–target characteristics (Günther, Dudschig, & Kaup, 2016; Hutchison, Balota, Cortese, & Watson, 2008; Jones & Golonka, 2012; Jones & Mewhort, 2007; Mandera, Keuleers, & Brysbaert, 2017). The idea is that relatedness is a continuous variable, rather than a dichotomous on/off switch, and this continuity should be reflected in the priming effect: some word pairs should consistently show a larger priming effect than others. The goal of the cited studies is to capture that variability using predictors like association strength, feature overlap, latent semantic analysis cosines, and the like. These predictors could, in turn, be considered as the organizing principles of semantic memory.

The success of such an approach critically depends on a frequently overlooked aspect, namely, the consistency of item-level priming effects. These priming effects can be estimated by subtracting the average response time to each target in the related condition (e.g., $\overline{RT}$ [*cat–dog*]) from the corresponding average response time in the unrelated condition (e.g., $\overline{RT}$ [*car–dog*]), thereby collapsing across participants. Crucially, it is often tacitly assumed that a particular prime yields a fairly *stable* advantage when it comes to processing a related target. As we mentioned before, some item-level priming effects are on average larger than others, such that, for instance, *cat* primes *dog* to a greater extent than does *animal*. However, these effects tend to vary considerably across participants (Heyman, Hutchison, & Storms, 2016a; Hutchison et al., 2008). For example, the reliability estimates for item-level priming effects obtained in the Semantic Priming Project (Hutchison et al., 2013) ranged from .08 to .35 (see Heyman, Hutchison, & Storms, 2016a, for more details). As such, the explanatory power of relatedness predictors is limited, because it is meaningless to predict noise.

By definition, item-level priming effects are the result of subtracting the average response time to a target in the related

[1] We use the term "semantic" to refer to semantic and/or associative priming.

✉ Tom Heyman
tom.heyman@kuleuven.be

[1] Department of Experimental Psychology, University of Leuven, Tiensestraat 102, 3000 Leuven, Belgium

[2] Montana State University, Bozeman, MT, USA

condition from the average response time to that target in the unrelated condition. Given that word recognition response times are not perfectly reliable (e.g., Brysbaert, Stevens, Mandera, & Keuleers, 2016), it follows that the *difference* between two response times is usually even less reliable (Guilford, 1954). That said, reliability estimates as low as .08 should be alarming, especially if one is interested in predicting item-level priming effects.

So, the goal of the present study was to obtain a precise estimate of the reliability of priming effects. Note that we will consider only *item-level* priming effects, not *person-level* priming effects. The split-half and test–retest reliability of the latter, assessed by collapsing across items instead of participants, have been examined in a number of studies already (Stolz, Besner, & Carr, 2005; Tan & Yap, 2016; Yap, Hutchison, & Tan, 2017). Critically, the reliabilities of person- and item-level priming effects are not necessarily related to one another. One concerns the stability of an item characteristic, the other the stability of an individual trait (see Heyman, Hutchison, & Storms, 2016a, for further discussion). To be clear, the present study's focus was also not on the *average* semantic priming effect (i.e., the grand mean). Indeed, it is well-established that the net priming effect after collapsing across participants *and* items is significantly greater than zero. Instead, the objective was to calculate a priming effect for *each item separately*, and to assess the reliability of these estimates.

The quite low reliability estimates of item-level priming effects observed in previous studies could partly be explained by interindividual variability in the degree of prime–target relatedness. Certain words might not yield priming effects for some participants, because the words are simply unknown to these participants. For instance, the low-frequency and fairly unfamiliar term *titmouse*, a small songbird, may not universally prime the target *bird*. Along the same lines, it is conceivable that two words are not connected (to the same extent) in every individual. There will be a strong link between *sprouts* and *disgusting* in some people's mind, whereas others may instead relate *sprouts* to *delicious*. More generally speaking, one could argue that each individual's semantic memory develops in its own unique way, which entails that item-level priming effects will always vary to some degree. Moreover, it is plausible that priming effects are not stable *within* subjects. The priming effect for, say, *cat–dog*, at time T_1 may be very different from the effect at time T_2. Temporary attention lapses, variability in the execution of motor responses, and the embedding in the experimental context (e.g., in the beginning of the experiment vs. at the end, or coincidentally preceded by a related prime–target pair, such as *snake–bite*) could all result in noisy priming effects.

In this study, we sought to gauge the potential impact of such inter- and intra-individual variability on semantic priming effects. More concretely, we assessed the test–retest (and split-half) reliability of item-level priming effects using classic psychometric methods (Lord, Novick, & Birnbaum, 1968). To do so, one needs to present the same target four times (i.e., twice in the related condition, *cat–dog*, and twice in the unrelated condition, *car–dog*). Typically, most semantic priming studies have avoided repeating items (but see, e.g., Durgunoğlu, 1988), because target responding might be influenced by episodic memory, which could potentially weaken semantic priming effects (McNamara, 2005). Nevertheless, adopting a test–retest design with sufficient spacing over time should mitigate such concerns. Concretely, we obtained item-level priming effects from every individual participant in two experimental sessions, separated in time by four weeks. We predicted that items with a relatively large priming effect at T_1 should also show a large priming effect at T_2. Given that Heyman, Hutchison, and Storms (2016a) found higher consistency across participants when the stimulus onset asynchrony (henceforth, SOA) was short (i.e., 200 ms) rather than long (i.e., 1,200 ms), we focused on conditions that should minimize the use of strategies (short SOA and low relatedness proportion—henceforth, RP). This study and others have suggested that there is at least some consistency. However, we did not have any clear, a priori expectations about the magnitude of test–retest reliability, since we measured item-level priming effects *for every individual participant*. In that sense, the present study was mostly exploratory.

Method

Pre-registration

We pre-registered the experiment before the data collection began, on the Open Science Framework (henceforth, OSF; see https://osf.io/7qgub/). The pre-registration contains a description of the research question, the sample size determination, the stimulus selection method, the stimulus material itself, the procedure of the experiment including the computer code, and an analysis plan.[2] We performed the data collection and analyses as described in the pre-registration, unless otherwise stated.

Participants

Fifty participants (42 women, eight men, mean age = 20 years) from the psychology department's participant pool at the University of Leuven took part in the experiment, in return for €14 or course credit. Forty participants completed both

[2] Due to an oversight, the analysis plan was not included in the initial pre-registration of the project. Unfortunately, this was only discovered after data collection for the first phase had been completed, so a new pre-registration of the project was created (see https://osf.io/c8sku/). Note, though, that the analysis plan was written before the data collection started and that no changes were made.

sessions (seven men, 33 women, mean age = 21 years). All participants were native Dutch speakers and reported no reading difficulties (e.g., dyslexia). The Social and Societal Ethics Committee of the University of Leuven approved the study (file number: G-2015 08 35), and participants provided written informed consent before the start of each session.

Materials[3]

The entire experiment consisted of two sessions, separated by four weeks. In each session, participants saw every target in both the related and the unrelated condition (see Table 1 for an illustration). Sessions comprised two blocks such that, in the first block, participants got half of the critical targets in the related condition and the other half in the unrelated condition. In the second block, the conditions were reversed. To determine the exact number of critical items, we conducted a power analysis using G*Power 3.1 (Faul, Erdfelder, Buchner, & Lang, 2009): Assuming a medium effect size of $|\rho| = .30$, two-tailed, $\alpha = .05$, and a power of .95, we needed a sample size of 134.

There were two lists, A and B, and participants were randomly assigned to one of the lists on the basis of participant number (i.e., odd numbers got List A, even numbers List B[4]). All primes appeared once per session (so, twice in total) and all targets appeared twice per session (so, four times in total). Sessions 1 and 2 were completely similar, except that the block order was reversed. The reason for creating two lists was to avoid repeating primes within a session. Furthermore, if we only focused on Block 1 of Session 1, we would have a typical priming design, in the sense that no stimulus had been presented more than once at that juncture. This allowed us also to estimate the split-half reliability of item-level priming effects in the same manner as Heyman, Hutchison, and Storms (2016a) and Hutchison et al. (2008).

Participants performed a lexical decision task under conditions that supposedly would minimize the use of strategies. More specifically, the RP was .25, the SOA 200 ms, and the nonword ratio (henceforth, NWR) .50. Given these task characteristics and the number of critical targets (i.e., 134), each block consisted of 67 critical related prime–target pairs, 67 critical unrelated prime–target pairs, 134 unrelated word–word pairs, and 201 word–nonword pairs. The filler targets (and primes) were not repeated within a session, but the exact same set of stimuli was used in the second session.

[3] The Materials & Procedure component of the project on the OSF describes the stimulus selection procedure in more detail.

[4] It was stipulated in the pre-registration that we would continue gathering data until 40 participants had completed both sessions. Because of a slightly unbalanced drop-out rate, 21 of these participants got List A, and 19 got List B.

Table 1. Simplified illustration of the study's design

	List A		List B	
	Prime	Target	Prime	Target
Block 1 Session 1	cat	dog	car	dog
	car	drive	cat	drive
	color	table	furniture	table
	furniture	blue	color	blue
Block 2 Session 1	vehicle	dog	bark	dog
	bark	drive	vehicle	drive
	chair	table	red	table
	red	blue	chair	blue
Block 1 Session 2	vehicle	dog	bark	dog
	bark	drive	vehicle	drive
	chair	table	red	table
	red	blue	chair	blue
Block 2 Session 2	cat	dog	car	dog
	car	drive	cat	drive
	color	table	furniture	table
	furniture	blue	color	blue

The original stimuli were in Dutch

Critical items came in the form of triplets such as *cat–bark–dog*, comprising a target (*dog*) and two semantically related primes (*cat* and *bark*). The 134 critical triplets were split into two sets of 134 related prime–target pairs. One set was featured in Block 1 of Session 1, the other in Block 2 of Session 1. For the first block, a typical counterbalancing procedure was applied: Half of the targets were preceded by their related prime in List A, whereas in List B they were preceded by an unrelated prime, and vice versa. The unrelated prime–target pairs were formed by recombining primes and targets (see Table 1 for an example). The same was true for the second block, except that the targets preceded by a related prime in Block 1 were preceded by an unrelated prime in Block 2, and vice versa.

As an illustration of the resulting item-level priming effects, consider the target *dog* in Table 1. Participants receiving List A saw the related prime *cat* before the target *dog* in Block 1 of Session 1, whereas participants with List B saw the unrelated prime *car* at this point. Hence, one can calculate an item-level priming effect (i.e., $\bar{RT}$ [*car–dog*] – $\bar{RT}$ [*cat–dog*]), and the same holds for all other critical targets in Block 1 of Session 1. We chose not to present the same primes in the second block of the session, to avoid repetition effects. Introducing a different set of primes in Block 2 did alter the listwise item-level priming effects. That is, the priming effect for *dog* in List A involved a comparison of *vehicle–dog* versus *cat–dog*, whereas in List B the comparison was *car–dog* versus *bark–dog*.

To make sure that the primes were actually related to their respective targets, a similarity metric based on the Dutch Word Association database was calculated (i.e., the random-walk spreading-activation measure described in De Deyne, Navarro, Perfors, & Storms, 2016). All targets were more closely related to their related than to their unrelated prime (see Fig. 1 for the frequency distribution of the difference scores), and the related pairs scored significantly higher on the relatedness metric than did the unrelated pairs [$t(133) = 32.54$, $p < .001$, $BF_{10} > 100$, and $t(133) = 36.19$, $p < .001$, $BF_{10} > 100$, for Lists A and B, respectively[5]]. Note that this result was not caused by related primes being more orthographically similar to the targets than were the unrelated primes. Indeed, the orthographic overlap (expressed in terms of Levenshtein distance) between the related prime–target pairs was similar to the orthographic overlap between the unrelated prime–target pairs [$t(133) = 0.47$, $p = .638$, $BF_{01} = 9.34$, and $t(133) = 0.94$, $p = .348$, $BF_{01} = 6.75$, for Lists A and B, respectively]. Furthermore, the stimulus set was as diverse as possible, including various types of prime–target pairs, such as synonyms (e.g., *gorgeous–pretty*), antonyms (e.g., *naughty–good*), category coordinates (e.g., *skirt–blouse*), subordinates (e.g., *animal–giraffe*), supraordinates (e.g., *chair–furniture*), property relations (e.g., *crow–black*), and (phrasal) associates (e.g., *cause–effect*). Selecting such a diverse sample increases generalizability and should help evade restriction-of-range issues. The rationale is that unreliability in terms of item-level priming effects might actually arise because the particular pairs are merely too similar. If we were to use only synonyms, for instance, the variability in item-level priming effects might be limited, which would translate into low reliability estimates. Indeed, the split-half and test–retest reliability estimates are correlations, so the same principles apply. As such, one would want to sample from the full relatedness range in order to avoid attenuating the resulting correlation/reliability coefficients.

Another 268 related word pairs were selected and subsequently recombined to form the unrelated fillers. A random half of the filler pairs appeared in Block 1 of Session 1 and the other half in Block 2 of Session 1 (this was the case for both Lists A and B). We then used the 402 targets (i.e., 134 critical + 268 filler stimuli) as the input to create nonwords. For each target, we selected a nonword from the Dutch Lexicon Project (Keuleers, Diependaele, & Brysbaert, 2010) that matched it in terms of length (i.e., number of characters) and orthographic typicality (i.e., the average Levenshtein distance of a stimulus to its 20 closest orthographic neighbors). Each nonword was preceded by a prime word, and again, a random half of these word–nonword pairs were shown in Block 1 of Session 1 and the other half in Block 2 of Session 1 (for both Lists A and B). Finally, an additional 20 unrelated word–word pairs and 20 word–nonword pairs were generated for the practice phase that immediately preceded each experimental session.

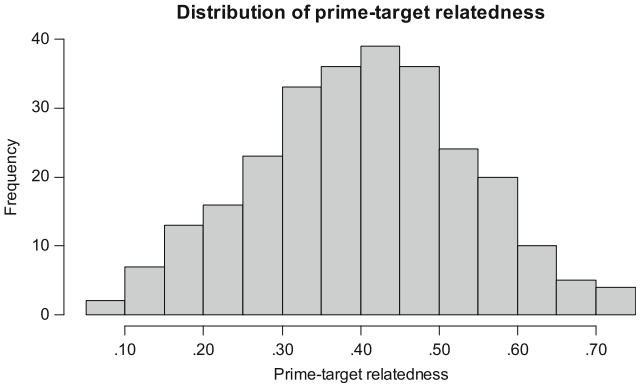

Fig. 1. Frequency distribution of prime–target relatedness. Values were obtained by subtracting the random-walk-based similarity (see De Deyne et al., 2016) for each unrelated prime–target pair from the corresponding similarity between the related prime and target. All values were positive, indicating that the related prime–target pairs were always considered to be more similar than their unrelated prime–target counterparts. The figure shows the distribution collapsed across Lists A and B, so $N = 268$

Procedure

In essence, participants did the same experiment twice. That is, the two sessions, separated by four weeks, were identical, except that the block order was reversed. The order of the trials within a block was completely random in both sessions. On every trial, participants first saw an uppercase prime, shown for 150 ms, followed by a blank screen, presented for 50 ms. Then, a lowercase target appeared, which was either a word or a nonword. Participants were asked to pay attention to the uppercase word and to then indicate as quickly and accurately as possible whether or not the lowercase letter string was an existing Dutch word. To respond, they had to press the left arrow for "word" or the right arrow for "nonword." The intertrial interval, during which a blank screen was shown, lasted 500 ms.

After signing the informed consent form, participants received instructions on how to perform the task. The experimenter first explained the procedure, after which the instructions also appeared on the computer screen. Each session started with a practice phase, followed by the experimental phase, which consisted of two blocks, both further subdivided into three parts. After each part, participants could take a self-paced break. The experiment was run on a Dell Pentium 4 with a 17.3-in. CRT monitor using Psychopy (Peirce, 2007). Each session took approximately 30 min.

[5] The abbreviation BF, short for "Bayes factor," has a subscript 10 indicating the relative plausibility of the data under the alternative hypothesis, $P(D \mid H_1)$, versus under the null hypothesis, $P(D \mid H_0)$. Throughout the text, we will use two subscripts 10 or 01 to indicate which term, $P(D \mid H_1)$ or $P(D \mid H_0)$, appears in the numerator and which in the denominator. The idea is to always present Bayes factors larger than 1, to facilitate interpretation of the results. Unless noted otherwise, we used the BayesFactor package (Morey & Rouder, 2015) and its default priors to calculate the Bayes factors.

Results

The present section is structured as follows. First, we report a number of "sanity checks" to verify that the methodology we employed was able to elicit semantic priming effects. This was indeed the case, which allowed us to then examine the test–retest and split-half reliabilities of the item-level priming effects. All analyses were conducted on the complete dataset, featuring the 134 crucial items, and on a dataset from which errors and outliers were removed. Outliers were defined per Participant × Session × Block combination, such that response times more extreme than three *SD*s above or below the participant-specific conditional mean were excluded (as had been the case in Heyman, Hutchison, & Storms, 2016a). This led to the exclusion of 3.93% of the data due to errors, and another 1.86% due to outliers. In general, the cleaned-up data yielded (slightly) better reliability estimates, so here we mainly focus on these results (also, because it is customary in the literature to perform some sort of data trimming). Similarly, most analyses were conducted on standardized response times, for these tend to improve the reliability of priming effects as compared to the raw response times (Heyman, Hutchison, & Storms, 2016a; Hutchison et al., 2008). Contrary to the analysis plan, which specified that outlier detection and *z*-transformation of the response times would occur per Participant × Session condition, we made a further subdivision by block. The rationale was that practice effects and the recurrence of the critical targets might affect the baseline response times in Block 2. To correct for this "block bias," we deemed it necessary to standardize per Participant × Session × Block combination. All analyses were carried out in R (version 3.3.1; R Development Core Team, 2016), and the analysis script, along with the raw data, is available at the OSF (see https://osf.io/y79fv/).

Exploratory and prerequisite analyses

First, item-level priming effects for the 134 crucial items were calculated for each Participant × Session combination separately. Averaging across items then yielded a person-level priming effect per session (see Table 2). The resulting priming effects, based on the *z*-transformed response times, were then subjected to a one-sample *t* test [$t(39) = 5.49$, $p < .001$, $BF_{10} > 100$, and $t(39) = 6.41$, $p < .001$, $BF_{10} > 100$, for Sessions 1 and 2, respectively]. Analogously, item-level priming effects for each Session × List condition were calculated by collapsing across participants, and these effects were in turn subjected to one-sample *t* tests [for List A in Session 1, $t(133) = 3.69$, $p < .001$, $BF_{10} = 54.93$; for List A in Session 2, $t(133) = 2.68$, $p = .008$, $BF_{10} = 2.97$; for List B in Session 1, $t(133) = 5.21$, $p < .001$, $BF_{10} > 100$; for List B in Session 2, $t(133) = 5.23$, $p < .001$, $BF_{10} > 100$]. Taken together, these results show that the applied methodology can in fact capture semantic priming. The magnitude of the effect is, on average, rather small, though. This is to be expected, given that the employed task characteristics should prevent strategic processes from boosting the overall effect. Moreover, the sample of prime–target pairs was deliberately heterogeneous in an attempt to cover the entire relatedness range. As a result, the average priming effect should be smaller than in studies using only strongly related word pairs (see Hutchison et al., 2013, for similar findings).

Table 2. Average error percentages, response times (RTs), and priming effects per session with the standard deviation in parentheses

Variable	Session 1	Session 2
Person-Level		
Unrelated % errors	4.38 (2.48)	4.48 (2.93)
Related % errors	3.41 (2.14)	3.43 (2.44)
Unrelated RT (ms)	629 (72)	559 (65)
Related RT (ms)	615 (69)	547 (63)
Priming RT (ms)	15 (20)	12 (13)
Priming zRT	0.13 (0.15)	0.12 (0.12)
Item-Level		
Unrelated % errors	4.38 (9.77)	4.48 (8.50)
Related % errors	3.40 (6.81)	3.40 (6.34)
Unrelated RT (ms)	633 (65)	561 (47)
Related RT (ms)	619 (65)	549 (42)
Priming RT (ms)	16 (89)	13 (56)
Priming zRT	0.13 (0.35)	0.13 (0.37)

Reliability of item-level priming effects

First, we examined whether participants' item-level priming effects from Session 1 correlated with those from Session 2. Figure 2 shows the distribution of the 40 test–retest

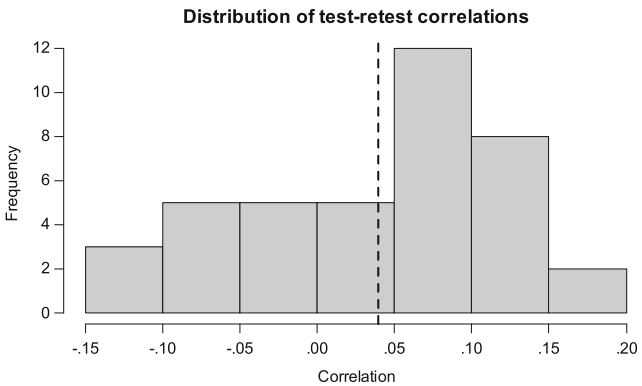

Fig. 2. Frequency distribution of the test–retest correlations. The dotted line indicates the average across 40 participants

correlations (i.e., one for every participant). As can be seen, most correlations were dishearteningly low. Only two of them significantly differed from zero (*p*s of .041 and .047), and a Bayesian test on the correlations (from Wetzels & Wagenmakers, 2012) always supported the null hypothesis to some degree (all BF_{01}s > 1.72).

Next, the average item-level priming effects per session were calculated by collapsing over participants. Because participants were assigned to one of two different lists, item-level priming effects for the 134 crucial targets could differ (e.g., $z\bar{R}T$ [*vehicle–dog*] – $z\bar{R}T$ [*cat–dog*] ≠ $z\bar{R}T$ [*car–dog*] – $z\bar{R}T$ [*bark–dog*]). Therefore, the priming effects were separated by list. Correlating these priming effects over sessions then provided the following test–retest reliability estimates: .29 for List A [$t(132) = 3.49$, $p < .001$, $BF_{10} = 22.61$] and .14 for List B [$t(132) = 1.61$, $p = .111$, $BF_{01} = 4.11$].[6]

Finally, we evaluated the split-half reliabilities by focusing only on Block 1 of Session 1. This is the "traditional" way of assessing the reliability of item-level priming effects, because at that point (i.e., the end of Block 1 in Session 1) participants had seen every prime and target just once (e.g., the target *dog* preceded by either the prime *cat*, for participants who were assigned to List A, or the prime *car*, for the participants who got List B). The participants receiving List A were split in two random halves, and so were the participants receiving List B. In a next step, average item-level priming effects were calculated for each half separately, which were subsequently correlated with one another. Applying the Spearman–Brown formula on the resulting correlations then gave one estimate of the split-half reliability. This procedure was repeated for 10,000 random halves, each yielding a reliability estimate. Averaging across those 10,000 estimates gave us a split-half reliability of .04.

The latter result is very surprising in light of previous studies (Heyman, Hutchison, & Storms, 2016a; Hutchison et al., 2008). Originally we planned to predict the reliability if we were to double the number of participants, but that seemed pointless, given the extremely low split-half reliability estimate. Instead, we conducted a number of nonplanned exploratory analyses to further investigate this unexpected result.

One potential implication of the present findings is that there were no meaningful differences between the selected items in terms of the priming effects they produced. Put differently, all items yielded priming effects that, on average, were very similar to one another. If true, this would imply that attempts to predict such item-level priming effects are futile and come down to explaining noise. However, as is shown in Fig. 1, prime–target relatedness seems to approximate a Gaussian distribution, meaning that few pairs are strongly or weakly related. Any (hypothetical) stable differences between these items might be overshadowed by the multitude of moderately related pairs eliciting inconsistent priming effects. To examine this possibility, we selected the 25% most strongly and 25% most weakly related items, on the basis of De Deyne et al.'s (2016) similarity metric. The split-half reliability, obtained using the same procedure described above, did increase to .12. That said, it still falls short of the estimates reported elsewhere (Heyman, Hutchison, & Storms, 2016a; Hutchison et al., 2008).

Besides a theoretically motivated attempt to boost reliability, we also tried a "kitchen sink" approach, which involved randomly picking 75% of the critical items. A thousand such random item samples were drawn, and for each subset the split-half reliability was estimated. This procedure resulted in a maximal reliability of .23 (see the left panel of Fig. 3 for a distribution of the estimates). In addition, we tested whether selecting only 75% of the *participants* would sometimes improve the consistency of the item-level priming effects, since it is possible that some participants performed the task very poorly and thereby muddled the general results. Again, we calculated the split-half reliabilities for 1,000 random samples, using all 134 items. The results revealed that the reliability estimate maximally increased to .25 (see the right panel of Fig. 3). Taken together, the reliabilities of some subsets seemed to be slightly better, yet still fairly poor. Even so, these appear to be outliers, given the many subsets with reliability estimates that were considerably lower.

Discussion

The present study paints a very sobering picture when it comes to the reliability of item-level priming effects. Test–retest, and especially split-half, reliability estimates proved extremely low, suggesting that attempts to predict such priming effects are unlikely to be successful. This might seem surprising, for two reasons: (a) Two previous studies have shown more reliable item-level priming effects (Heyman, Hutchison, & Storms, 2016a; Hutchison et al., 2008), and (b) other studies have found significant relations between priming and various semantic relatedness variables (Günther et al., 2016; Hutchison et al., 2008; Jones & Golonka, 2012; Jones & Mewhort, 2007; Mandera et al., 2017). How can we explain these ostensible discrepancies?

First of all, in the present study we rigorously controlled the orthographic overlap between the primes and targets. That is, the related prime–target pairs were closely matched with the unrelated pairs in terms of the Levenshtein distance between the prime and target. In many other priming studies, some

[6] Though this is not mentioned in the analysis plan, we also calculated the test–retest reliabilities of the related and unrelated response times as such. These results showed estimates of .72 (List A, related condition), .50 (List B, related condition), .67 (List A, unrelated condition), and .65 (List B, unrelated condition).

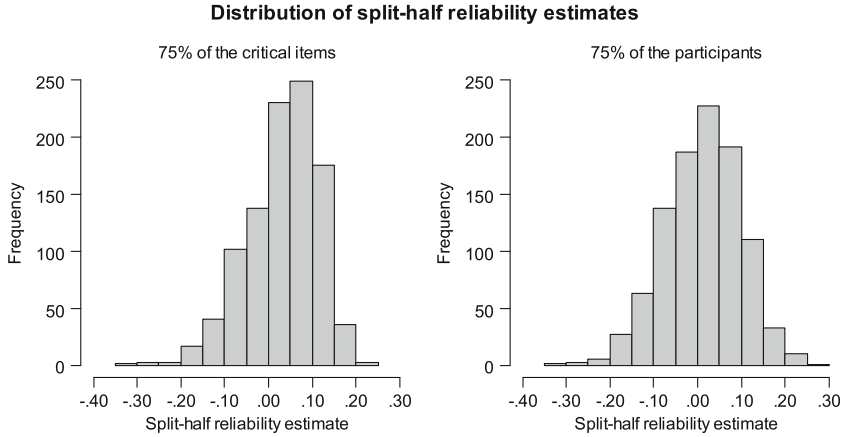

Fig. 3. Frequency distributions of split-half reliability estimates using 75% of the critical items (left) or 75% of the participants (right)

items were not only semantically, but also orthographically and morphologically related (e.g., *abnormal–normal*). According to lexical decision data from Rastle and colleagues (Rastle, Davis, Marslen-Wilson, & Tyler, 2000), the priming effect for suchlike pairs increases by 10–20 ms on average, when compared with prime–target combinations that are only semantically related. Such a morphological/orthographic *boost* might in turn inflate the overall reliability of item-level priming effects. The present results might therefore indicate that pure(r) semantic priming effects are (even) less reliable, though this assertion is admittedly speculative.

Another difference from the studies cited above is that the task conditions we employed should have limited the use of strategies. It could be that the obtained, presumably automatic, priming effects were (even) less reliable than more strategic priming effects. Although this runs counter to the claims we made previously (Heyman, Hutchison, & Storms, 2016a), it would support Stolz et al.'s (2005) hypothesis that activity in semantic memory is very noisy. To address this possibility, we reanalyzed the data of three recently published priming studies with task characteristics that are also thought to curtail strategic processing: the low-relatedness condition of de Wit and Kinoshita (2015); Experiment 3 of Heyman, De Deyne, Hutchison, and Storms (2015); and the semantic priming condition of Tan and Yap (2016).[7] The former study was very similar to the present experiment, in that it used a standard lexical decision task with a short SOA (240 ms), unmasked primes, and a low relatedness proportion (.25).[8] Experiment 3 of Heyman et al. (2015) involved a continuous lexical decision task, which has been argued to reduce strategic effects, since it requires responses to both the primes and targets (Shelton & Martin, 1992). Finally, the primes in Tan and Yap's study were masked, again decreasing the likelihood of strategic influences. Even though all three studies ought to have limited the use of strategies, the split-half reliability estimates of the item-level priming effects, obtained via the method explained above, varied from .03 (de Wit & Kinoshita, 2015) to .15 (Heyman et al., 2015) to .70 (Tan & Yap, 2016). The wide range may seem surprising at first, but a clearer picture emerges if we consider the sample sizes. The estimates of the item-level priming effects from Tan and Yap are based on 240 participants, about six to eight times more than in the other studies discussed here (i.e., 29 participants in the low-relatedness condition of de Wit and Kinoshita's, 2015, study, and 40 participants in Exp. 3 of Heyman et al., 2015). That is not to say that sample size is the only factor determining the reliability of priming effects, but it certainly is a *big* factor.

To illustrate this assertion, we conducted a follow-up analysis on Tan and Yap's (2016) data. That is, we selected random subsets of 40 participants, the same number of participants as in the present experiment and in Heyman et al.'s (2015), for which we then calculated the split-half reliability. The resulting estimate dropped from .70 (based on all participants) to a paltry .28 (based on 40 participants).[9] Note that these findings also imply that item-level priming effects are not by definition unreliable, thus nuancing Stolz et al.' (2005) claim that semantic activation is noisy and uncoordinated. It appears that one just needs sufficiently large samples in order to obtain stable estimates.

The question then becomes how to determine sample size. One option would be to base it on previous studies like Tan and Yap's (2016). However, it is possible that reliability coefficients vary as a result of using different item samples and

[7] The studies from de Wit and Kinoshita (2015) and Tan and Yap (2016) did not directly focus on item-level priming effects.

[8] De Wit and Kinoshita (2015) actually manipulated relatedness proportion between participants (with two levels, .25 and .75). The data from the high-relatedness proportion condition were not considered here.

[9] One could also apply Spearman–Brown's prediction formula to derive this figure. Although it is typically used to predict the reliability with increasing test length (i.e., number of participants, in our case), one could also do the opposite. Filling in the obtained reliability estimate including all participants (.70) and the factor by which the sample size was "increased" (40/240 = 0.17) gives us (.70 × 0.17)/(1 – 0.83 × .70) = .28. This is exactly equal to the average split-half reliability estimate obtained for subsamples of Tan and Yap's (2016) data.

task conditions. Another, potentially safer alternative would be to conduct an initial study with sample size k' (say, 50 participants) and calculate the corresponding reliability of the item-level priming effects $\rho_{k'}$. On the basis of the present results, we can assume that the resulting reliability estimate would not be high enough. Yet, we can derive the sample size k that would be necessary to achieve the desired reliability ρ_k as follows (see Lord et al., 1968):

$$k = k' \frac{\rho_k \left(1-\rho_{k'}\right)}{\rho_{k'} \left(1-\rho_k\right)}.$$

For example, suppose that the aspired-to reliability coefficient were .80, and that one had obtained a reliability of .40 with 50 participants. Plugging in those values in the formula above suggests that one would need 300 total participants—so, 250 in addition to the 50 initial participants. Importantly, we assume here that every critical target is presented equally often in the related and unrelated conditions, so that 150 participants would receive *cat–dog* and the other 150 would get *car–dog*. Note that if one wanted to obtain reliable item-level priming effects under different circumstances (e.g., short vs. long SOA), it would be advisable to follow the outlined procedure for each condition separately. In any case, once all the data have been gathered, one ought to calculate the reliability again, since the formula described above only provides an estimate based on a number of assumptions (e.g., that all additional participants provide equally reliable data) that might be violated.

To further illustrate this approach, we once more used Tan and Yap's (2016) data. Let's assume that the reliability of the item-level priming effects turned out to be .28 after an initial 40 participants took part in the experiment. Figure 4 shows the sample sizes that would be required in this case to reach a certain level of reliability. As can be seen, increasingly larger sample sizes are needed to achieve reliability coefficients toward the upper end of the spectrum. In practice, one would need to weigh the costs of having to gather a considerable amount of data against the benefit of obtaining greatly better priming effect estimates.

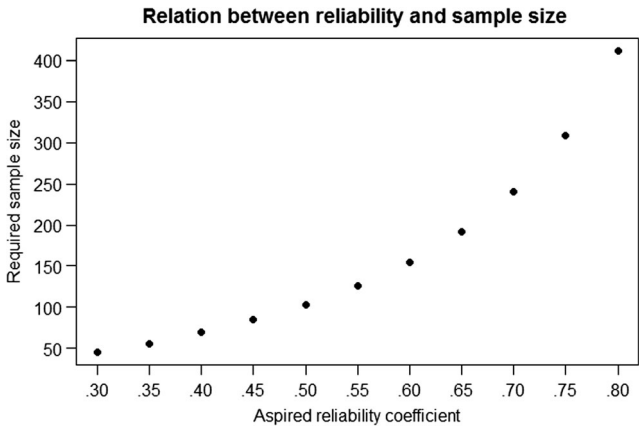

Fig. 4. Sample sizes (on the *y*-axis) required to reach certain levels of reliability (on the *x*-axis), assuming an initial reliability coefficient of .28 with a sample size of 40, as in Tan and Yap (2016), based on a reformulation of the Spearman–Brown prediction formula. Note that as the reliability coefficient approaches 1, the sample size further increases disproportionately (e.g., 1,955 participants would be needed to achieve a reliability of .95)

In the literature such large samples have typically been lacking, unfortunately. Those studies that have collected data from several hundreds of participants (e.g., Hutchison et al., 2013) usually had designs with multiple variables manipulated within items. Consequently, the number of observations per cell can become relatively small, which results in noisy item-level priming effects when one considers each condition separately. One could, of course, collapse the data across conditions in order to obtain more reliable, generic priming effects, but that is not ideal and might defeat the purpose of the study (e.g., comparing the item-level priming effects when SOA is short vs. long). In sum, given the current state of the field, it is difficult to fully understand the potential impact of factors like morphological/orthographic similarity, SOA, and relatedness proportion. One might even argue that such a comparison is futile at this point, because, with the exception of Tan and Yap's (2016) complete dataset, the resulting item-level priming effects were too unreliable to begin with. Put differently, we might speculate about the role of, say, morphological/orthographic similarity (see above), but the data currently available do not warrant any strong conclusions one way or another.

Thus far, we have exclusively focused on item-level priming effects in the context of a lexical decision task. As a consequence, one might wonder whether the task itself is (partly) responsible for the lack of reliable estimates. Previous studies have provided evidence against this notion, though. That is, the same issue arises when using different paradigms, such as speeded naming (Heyman, Hutchison, & Storms, 2016a; Hutchison et al., 2008) or speeded word fragment completion (Heyman et al., 2015). Furthermore, a reanalysis of de Wit and Kinoshita's (2014) semantic categorization data (i.e., the low-relatedness-proportion condition of their Exp. 1) yielded item-level priming effects with a split-half reliability of merely .04. Taken together, all paradigms face the same problem, in that they often produce unreliable item-level priming effects. That said, pooling estimates from different tasks might be a fruitful approach. Not only would one effectively increase the sample size that way, but the technique might also allow researchers to filter out some task-specific noise.

Predicting related response times

Several of the studies that have tried to predict priming effects didn't actually use difference scores as the dependent variable. Instead, they predicted response times to the targets in the

related condition (aggregated across participants), possibly because these are more reliable, as such (i.e., the estimates typically vary from .60 to .80). One may wonder whether this could be a better approach, since it appears to circumvent the reliability issue that plagues difference scores (i.e., the item-level priming effects; see above). We will argue, though, that it faces less obvious, yet similar problems.

The critical issue here is that one needs one or more covariates in order to statistically control for *baseline* response times to the target. After all, there are many factors that determine target recognition times. Context is certainly one of them, but its relative contribution is generally limited. Thus, one needs to make sure that the relatedness variables aren't just word recognition predictors in disguise. The reality is that such variables usually correlate negatively with target response times. This is most obvious for prime–target co-occurrence frequency, which naturally depends on target frequency, a strong predictor of response times (i.e., high-frequency words tend to be recognized faster). So, in order to unequivocally establish the relation between any relatedness predictor and semantic priming, one should adequately control for the baseline target response time. If that is not the case, the relatedness variable(s) might actually explain variability between the targets in baseline response time, rather than providing a meaningful account of semantic priming.

There are two, potentially complementary, ways to address this concern. First of all, one could add a bunch of covariates that have been shown to predict word recognition times. The stumbling block is that no set of variables (yet) can explain *all* the systematic variability in baseline response times (e.g., Adelman, Marquis, Sabatos-DeVito, & Estes, 2013; Heyman, Van Akeren, Hutchison, & Storms, 2016b). Thus, it is simply not possible to rule out that the relatedness variable(s) merely predict target response times instead of semantic priming, even when controlling for word frequency, length, orthographic/phonological neighborhood density, age of acquisition, and the like.

A second option is to (also) include a measure of baseline response time as a covariate, obtained either from the same experiment (e.g., the response times to the targets in the unrelated condition) or from a norming study (e.g., Balota et al., 2007). However, this solution isn't ideal, either, again because of concerns about (un)reliability. As was demonstrated by Westfall and Yarkoni (2016), even moderately reliable covariates inflate Type I error rates when one seeks to establish incremental validity (e.g., "co-occurrence frequency predicts target response times in the related condition when controlling for baseline response times"). Hence, one needs very reliable baseline response times in order to draw meaningful conclusions about the predictive value of any relatedness variable. Unfortunately, this is usually not the case. Suppose, for instance, that we had predicted target response times in the related condition on the basis of the prime–target co-occurrence frequency and response times in the unrelated condition. The reliability estimate of the latter variable was .60, so we would have a fairly imprecise approximation of the baseline response time. The relatedness variable (i.e., prime–target co-occurrence frequency, in this example) could "pick up the slack" and explain some of the variability in the dependent variable that is actually associated with baseline word recognition. So the finding that a relatedness variable predicts target response times in the related condition when statistically controlling for baseline response times may be inconsequential when it comes to explaining semantic priming. It could reveal something real about priming, but it might also reflect baseline word recognition, or it might relate to both. The problem is that one can't disentangle these alternatives unless one were to boost the reliability of the baseline response times.

To illustrate this issue, we used Westfall and Yarkoni's (2016) application (http://jakewestfall.org/ivy/) to calculate what the Type I error rate would be for the present dataset, again looking only at Block 1 of Session 1, as to mimic the typical priming design. Let us assume that the reliability of the relatedness predictor in question—say, prime–target co-occurrence frequency—is .90, and that its true correlation with baseline response time is a moderate .30. Plugging in these values, together with the estimated true correlation between related and unrelated response times (.97), the reliability of the unrelated (baseline) response time (.60), and the sample size (134), yields a Type I error rate of .51. This rate drops as the true correlation between baseline response time and the relatedness variable goes to zero, but it is still .10 for a correlation as low as .10. In other words, the Type I error rate is often considerably higher than the typical .05 level, meaning that incremental validity claims are (very) error-prone. The problem with this approach is not that the dependent variable, response times to targets preceded by their related primes, is unreliable. Rather, the moderately reliable baseline response time covariate, as observed here and in other (norming) studies, opens the door for spurious relations between the dependent variable and relatedness predictors. Indeed, if we were to improve the reliability of the baseline response times to .95 by substantially increasing the number of participants, the Type I error rates for the examples noted above would shrink to .08 (from .51) and .05 (from .10). Using Tan and Yap's (2016) data, for instance, would yield Type I error rates of .06 and .05, respectively. However, if we again were to randomly select only 40 participants, error rates would rise markedly, to .21 and .07, respectively.

In sum, it seems that both approaches—predicting item-level priming effects or target response times in the related condition—face similar obstacles. Most estimates are just not reliable enough to truly allow us to understand how priming arises and what it tells us about semantic memory. In the first approach, predicting item-level priming, the issue is a lack of power. If measures of the dependent variable are

completely unreliable—as was, for instance, the case in the present study—it is impossible to establish meaningful relations with any relatedness predictor. The situation becomes less troubling as the reliability of item-level priming effects increases (as it did in Hutchison et al., 2008), but even with estimates hovering around .30–.40, subtle relations may go undetected. In the second approach, predicting target response times, power is less of a concern, but elevated Type I error rates are. Because estimates of the covariate (i.e., baseline response time) are typically only moderately reliable, it is difficult to establish meaningful relations with relatedness predictors. Luckily, these are not insurmountable problems. As was evidenced, for instance, by the reanalyses of Tan and Yap's (2016) data, a larger sample size can considerably improve the reliability of response time estimates. Thus, big data, in terms of items *and* participants, should be the future for semantic priming research (see also Stevens & Brysbaert, 2016, for the same message in the context of mixed-effect analyses of semantic priming).

Author note T.H. developed the study concept. A.B. gathered the data. T.H. and A.B. performed the data analysis and interpretation. T.H. drafted the manuscript, and K.A.H. and G.S. provided critical revisions. All authors approved the final version of the manuscript for submission. We thank Bianca de Wit and Melvin Yap for kindly providing their data, as well as Simon De Deyne for calculating the random-walk-based similarities. T.H. is a Postdoctoral Fellow of the Research Foundation–Flanders (FWO-Vlaanderen).

References

Adelman, J. S., Marquis, S. J., Sabatos-DeVito, M. G., & Estes, Z. (2013). The unexplained nature of reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1037–1053. doi:https://doi.org/10.1037/a0031829

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., … Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*, 445–459. https://doi.org/10.3758/BF03193014

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. *Journal of Experimental Psychology: Human Perception and Performance*, *42*, 441–458. https://doi.org/10.1037/xhp0000159

De Deyne, S., Navarro, D. J., Perfors, A., & Storms, G. (2016). Structure at every scale: A semantic network account of the similarities between unrelated concepts. *Journal of Experimental Psychology: General*, *145*, 1228–1254. https://doi.org/10.1037/xge0000192

de Wit, B., & Kinoshita, S. (2014). Relatedness proportion effects in semantic categorization: Reconsidering the automatic spreading activation process. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*, 1733–1744. https://doi.org/10.1037/xlm0000004

de Wit, B., & Kinoshita, S. (2015). An RT distribution analysis of relatedness proportion effects in lexical decision and semantic categorization reveals different mechanisms. *Memory & Cognition*, *43*, 99–110. https://doi.org/10.3758/s13421-014-0446-6

Durgunoğlu, A. Y. (1988). Repetition, semantic priming, and stimulus quality: Implications for the interactive-compensatory reading model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 590–603. https://doi.org/10.1037/0278-7393.14.4.590

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Guilford, J. P. (1954). Psychometric methods (2nd ed.). New York, NY: McGraw-Hill.

Günther, F., Dudschig, C., & Kaup, B. (2016). Latent semantic analysis cosines as a cognitive similarity measure: Evidence from priming studies. *Quarterly Journal of Experimental Psychology*, *69*, 626–653. https://doi.org/10.1080/17470218.2015.1038280

Heyman, T., De Deyne, S., Hutchison, K. A., & Storms, G. (2015). Using the speeded word fragment completion task to examine semantic priming. *Behavior Research Methods*, *47*, 580–606. https://doi.org/10.3758/s13428-014-0496-5

Heyman, T., Hutchison, K. A., & Storms, G. (2016a). Uncovering underlying processes of semantic priming by correlating item-level effects. *Psychonomic Bulletin & Review*, *23*, 540–547. https://doi.org/10.3758/s13423-015-0932-2

Heyman, T., Van Akeren, L., Hutchison, K. A., & Storms, G. (2016b). Filling the gaps: A speeded word fragment completion megastudy. *Behavior Research Methods*, *48*, 1508–1527. https://doi.org/10.3758/s13428-015-0663-3

Hutchison, K. A., Balota, D. A., Cortese, M. J., & Watson, J. M. (2008). Predicting semantic priming at the item level. *The Quarterly Journal of Experimental Psychology*, *61*, 1036–1066. https://doi.org/10.1080/17470210701438111

Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C.-S., … Buchanan, E. (2013). The semantic priming project. *Behavior Research Methods*, *45*, 1099–1114. https://doi.org/10.3758/s13428-012-0304-z

Jones, L. L., & Golonka, S. (2012). Different influences on lexical priming for integrative, thematic, and taxonomic relations. *Frontiers in Human Neuroscience*, *6*, 205. https://doi.org/10.3389/fnhum.2012.00205

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*, 1–37. https://doi.org/10.1037/0033-295X.114.1.1

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono- and disyllabic words and nonwords. *Frontiers in Psychology*, *1*, 174. https://doi.org/10.3389/fpsyg.2010.00174

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). Statistical theories of mental test scores. Oxford, UK: Addison-Wesley.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78. https://doi.org/10.1016/j.jml.2016.04.001

McNamara, T. P. (2005). Semantic priming: Perspectives from memory and word recognition. New York, NY: Psychology Press.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*, 227–234. https://doi.org/10.1037/h0031564

Morey, R. D., & Rouder, J. N. (2015). BayesFactor: Computation of Bayes factors for common designs (Version 0.9.12–2). Retrieved from https://CRAN.R-project.org/package=BayesFactor

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*, 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

R Development Core Team. (2016). R: A language and environment for statistical computing (Version 3.3.1). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from www.R-project.org

Rastle, K., Davis, M. H., Marslen-Wilson, W. D., & Tyler, L. K. (2000). Morphological and semantic effects in visual word recognition: A time-course study. *Language and Cognitive Processes*, *15*, 507–537. https://doi.org/10.1080/01690960050119689

Shelton, J. R., & Martin, R. C. (1992). How semantic is automatic semantic priming? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 1191–1210. https://doi.org/10.1037/0278-7393.18.6.1191

Stevens, M., & Brysbaert, M. (2016). *When do we have enough power in language research? Evidence from priming studies*. Unpublished manuscript. Retrieved from http://crr.ugent.be/papers/When%20do%20we%20have%20enough%20power%20in%20language%20research.pdf

Stolz, J. A., Besner, D., & Carr, T. H. (2005). Implications of measures of reliability for theories of priming: Activity in semantic memory is inherently noisy and uncoordinated. *Visual Cognition*, *12*, 284–336. https://doi.org/10.1080/13506280444000030

Tan, L. C., & Yap, M. J. (2016). Are individual differences in masked repetition and semantic priming reliable? *Visual Cognition*, *24*, 182–200. https://doi.org/10.1080/13506285.2016.1214201

Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PLOS ONE*, *11*, e0152719. https://doi.org/10.1371/journal.pone.0152719

Wetzels, R., & Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review*, *19*, 1057–1064. https://doi.org/10.3758/s13423-012-0295-x

Yap, M. J., Hutchison, K. A., & Tan, L. C. (2017). Individual differences in semantic priming performance: Insights from the Semantic Priming Project. In M. N. Jones (Ed.), Big data in cognitive science (pp. 203–226). New York, NY: Psychology Press.