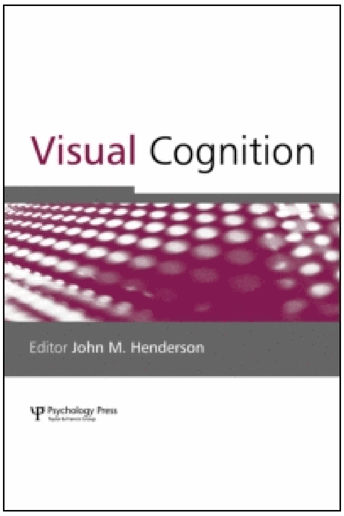

Visual Cognition
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713683696

Visual word recognition: On the reliability of repetition priming

Stephanie Waechter [a]; Jennifer A. Stolz [a];Derek Besner [a]
[a] Psychology Department, University of Waterloo, Ontario, Canada

First published on: 24 June 2009

VISUAL COGNITION, 2010, 18 (4), 537–558

Psychology Press
Taylor & Francis Group

Visual word recognition: On the reliability of repetition priming

Stephanie Waechter, Jennifer A. Stolz, and Derek Besner
Psychology Department, University of Waterloo, Ontario, Canada

Repetition priming is one of the most robust phenomena in cognitive psychology, but participants vary substantially on the amount of priming that they produce. The current experiments assessed the reliability of repetition priming within individuals. The results suggest that observed differences in the size of the repetition priming effect across participants are largely reliable and result primarily from systematic processes. We conclude that the unreliability of *semantic priming* observed by Stolz, Besner, and Carr (2005) is specific to uncoordinated processes in semantic memory, and that this unreliability does not generalize to other processes in visual word recognition. We consider the implications of these results for theories of automatic and controlled processes that contribute to priming. Finally, we emphasize the importance of reliability for researchers who use similar paradigms to study individual and group differences in cognition.

Keywords: Repetition priming; Statistical reliability; Repetition proportion; Visual word recognition.

The *reliability* of a measure refers to its consistency at the level of the individual. Reliability is an important psychometric property that should be evaluated in the development of any behavioural measure, in part because the reliability of a measure greatly constrains its usefulness. Low reliability reduces the chance of finding both significant correlations between measures and significant differences between groups.

Please address all correspondence to Stephanie Waechter, Department of Psychology, University of Waterloo, 200 University Ave. W., Waterloo, Ontario, Canada N2L 2G1. E-mail: swaechte@uwaterloo.ca

This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada to DB and JAS, and preliminary results were presented at the 18th annual meeting of the Canadian Society for Brain, Behaviour, and Cognitive Science in London, Ontario. We thank Keith A. Hutchison and two anonymous reviewers for their comments on an earlier version of this manuscript. We also thank Sandra C. Gregory, Colleen Merrifield, Zhenhua Wang, Regina Thurairatnam, and Melissa Chartrand for their help with data collection.

http://www.psypress.com/viscog DOI: 10.1080/13506280902868603

Kopriva and Shaw (1991) provided an interesting demonstration of the impact of low reliability on the power to detect differences between groups. Imagine that two populations actually differ by one standard deviation on a measure. A researcher recruits 25 participants from each population, and seeks to compare these two groups. When the reliability of the dependent measure is .40, the power to detect a significant difference ($\alpha = .05$) with a one-way ANOVA is only .50. When the reliability of the dependent measure is .80, however, the power to detect this same difference increases to .80. Using measures with low reliability can therefore substantially diminish the likelihood of detecting real differences between groups.

THE RELIABILITY OF PRIMING

We are concerned with the reliability of *priming*: The benefits in response time accrued when participants respond to a stimulus that is preceded by a related stimulus compared to an unrelated stimulus. Note that this definition is distinct from the term "priming" in some of the memory literature, where it denotes changes in accuracy in tasks such as word-stem completion. For the purposes of the current study, we will be using the term priming to refer specifically to response time differences.

Priming scores have been used as a tool to examine differences in implicit memory between Alzheimer's patients and controls (e.g., Balota, Black, & Cheney, 1992; Balota & Duchek, 1991; Nebes, Brady, & Huff, 1989) and between depressed individuals and controls (e.g., Danion et al., 1991). When significant between-group differences are not found, researchers often conclude that the group of interest does not suffer from deficits in implicit memory. However, the absence of between-group differences may instead simply reflect the unreliability of the dependent measure. Similarly, studies that attempt to correlate priming scores with other continuous measures may suffer the same problem. The absence of a correlation between two measures with low reliability does not necessarily mean that the two constructs are unrelated. Therefore, assessing the reliability of priming scores represents a crucial methodological issue (see Hutchison, Balota, Cortese, & Watson, 2008, and Madden, Pierce, & Allen, 1993, for similar discussions).

RELIABILITY OF SEMANTIC PRIMING

Stolz, Besner, and Carr (2005) examined the reliability of *semantic priming* scores. They did this by examining the correlations between participants' individual semantic priming scores for Block 1 and their individual semantic priming scores for Block 2. Blocks 1 and 2 contained separate lists of word-prime and word-target pairs that were matched across blocks on various characteristics, including strength of association. Therefore, Stolz et al. administered two different but comparable sets of test items and then correlated participant scores on the two, in a procedure similar to the calculation of test–retest reliability. They were primarily interested in whether or not semantic priming occurs in a stable, predictive, and consistent way at the level of the individual. For example, did participants who showed larger-than-average semantic priming scores in Block 1 also show comparably large priming effects in Block 2?

The results showed, at best, only modest reliability. In fact, the highest correlation in any condition was $r = .38$, indicating that priming scores in Block 1 predicted at most 14% of the variance in priming scores in Block 2. Also, although their experiment included *nine* different between-participants conditions (three relatedness proportions, RPs, crossed with three stimulus–onset asynchronies, SOAs) reliability was not significant in five of these conditions. These nonsignificant correlations ranged from $-.15$ to $+.10$. Therefore, researchers who use semantic priming as a tool to examine individual and group differences must carefully consider reliability when interpreting their results. An absence of group differences may simply reflect the low reliability of the dependent measure, rather than any real property of the groups being studied.

IMPLICATIONS FOR VISUAL WORD RECOGNITION

In addition to examining the implications for individual and group differences research, Stolz et al. (2005) also sought to explain *why* they observed such low reliability. Stolz et al. attributed the observed lack-of-reliability to activity in semantic memory, which they suggested is "inherently noisy and uncoordinated" (p. 328).

However, there is a potential alternative explanation. The observed unreliability of semantic priming scores need not reflect uncoordinated activity in semantic memory at all. Most contemporary theories of visual word recognition (e.g., Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; McClelland & Rumelhart, 1981; Perry, Zeigler, & Zorzi, 2007) distinguish between several presemantic levels of processing. For example, a recent extension of McClelland and Rumelhart's (1981) Interactive Activation model (Stolz & Besner, 1996) consists of a feature level, a letter level, a lexical level, and a semantic level. Excitatory activation in this model cascades forward from the feature level (where the features that make up letters are activated) to the letter level (where individual letters are activated) to the lexical level (where lexical representations are activated) and finally to the semantic level (where semantic representations are activated). Activation in this model also feeds back from the semantic level to the lexical level and from the lexical level to the letter level. Therefore, although activity in semantic memory may be inherently noisy, the observed unreliability of semantic priming scores could instead reflect noisy and uncoordinated processes at feature, letter, and/or lexical levels of processing.

How can we determine whether the unreliability of semantic priming is specific to processing in semantic memory? One approach is to examine the reliability of a different form of priming, one that is not as strongly dependent on activity in semantic memory.

THE RELIABILITY OF REPETITION PRIMING

Repetition priming can be assessed in terms of the benefits in RT and accuracy accrued when participants respond to a repeated stimulus compared with a unrelated stimulus. In the domain of visual word recognition, repetition priming scores reflect the decreases in RT (and errors) observed when participants respond to a word that is immediately preceded by the same word, rather than by a different, unrelated word (e.g., Scarborough, Cortese, & Scarborough, 1977). Repetition priming is widely regarded as stemming from either residual activation of the target word's entry in the lexicon (e.g., Morton, 1969), or from the retrieval of earlier episodes involving the target word (e.g., Jacoby, 1983). Though the debate over these two theories remains somewhat unresolved (see Bowers, 2000, and Tenpenny, 1995, for reviews), it is clear that repetition priming reflects either residual activation or facilitated retrieval of representations that are not merely semantic. Therefore, repetition priming represents a suitable tool for examining the reliability of lexical and sublexical processes in visual word recognition.

THE CURRENT STUDY

The purpose of the current study is twofold. The first purpose is to compare the reliability of repetition priming scores with the reliability of semantic priming scores, in order to determine whether the observed unreliability of semantic priming is specific to semantic-level processing or merely an artefact of the inherent unreliability of lower level processes in visual word recognition. If the pattern of reliability we observe for repetition priming scores is similar to the pattern observed by Stolz et al. (2005) for semantic priming scores, then lexical or prelexical level processes in visual word recognition must frequently occur in an uncoordinated and inconsistent manner. However, if repetition priming is more reliable under a broader range of conditions than is semantic priming, we can conclude that the unpredictability of semantic priming scores is likely due specifically to the unreliability of processes in semantic memory. In order to allow a direct comparison, the present conditions, materials, and task (lexical decision) were largely identical to those used by Stolz et al.

The second purpose of the current study is to examine the reliability of repetition priming scores across different conditions. What can we expect to happen to the reliability of repetition priming scores with varying RPs and SOAs? Stolz et al. (2005) found significant reliability in four conditions: At all three SOAs when the RP was .50, and at the longest SOA when RP was .75 (see Table 2 for a summary). Stolz et al. interpreted their findings in terms of a combination of two mechanisms. As RP increased from .25 to .50, the retrieval of prime episodes (e.g., Bodner & Masson, 2001) became increasingly systematic, resulting in significant reliability. When RP increased to .75, however, Stolz et al. posited that participants began to engage in *strategic* generation of potential target candidates. When the SOAs were short, these expectancies did not have time to succeed and semantic priming was not reliable. Stolz et al. concluded that it is the combination of automatic episodic retrieval and strategic target generation that determines the reliability of semantic priming.

As Stolz et al. (2005) explained, repetition priming at low RPs and short SOAs is likely to reflect more automatic and reflexive processing, whereas higher RPs combined with longer SOAs increase the likelihood that controlled, effortful, and strategic processes will be recruited. Therefore, the current experiment also explores the contributions of automatic and controlled processes to the reliability of repetition priming.

EXPERIMENT 1

Method

Participants. One hundred and forty-four undergraduate students from the University of Waterloo participated in this experiment in exchange for course credit. All participants spoke English as their first language and had normal or corrected-to-normal vision.

Design. Each experiment employed a 2 (Priming: Repeated vs. Unrelated) × 2 (Block: Block 1 vs. Block 2) within-participants design. The stimulus–onset asynchrony (SOA) between the prime and target was 200 ms in Experiment 1A, 350 ms in Experiment 1B, and 800 ms in Experiment 1C.

Stimuli. The 200 word prime-word target pairs and the 200 word prime-nonword target pairs from Stolz et al. (2005) were used in Experiments 1A, 1B, and 1C. New repeated word trials were constructed by replacing the original prime word with the target word (e.g., the original semantically related pair "east-west" was changed to "west-west"). No other items or pairs were altered.

List construction was largely identical to that used by Stolz et al. (2005). In each experiment, we analysed data from 25 repeated word-prime word-target trials and 25 unrelated word-prime word-target trials from each block. Buffer trials, which were not analyzed, were used to create the desired priming proportions. In Experiments 1A, 1B, and 1C, 50 unrelated word-prime word-target buffer pairs were present in each block. Buffer trials were counterbalanced so that all prime-target pairs served as buffer and critical trial pairs across participants. Each block then contained 75 unrelated word-prime word-target pairs, 25 repeated word-prime word-target pairs, and 100 word prime-nonword target pairs, to create an RP of .25. The assignment of pairs to Block 1 or Block 2 was also counterbalanced across participants.

An additional list of 20 practice trials (10 word prime-word target and 10 word-prime nonword target pairs) was constructed to roughly match the proportions used in the rest of the experiment (e.g., RP = .2 in Experiment 1A). The practice trials were administered to each participant before the experiment trials.

Procedure. Participants were tested individually, seated approximately 50 cm from a 15-inch computer monitor. Task instructions were displayed on the monitor and were also relayed verbally.

Stimulus display and response collection were controlled by E-Prime software (Psychology Software Tools). Practice trials were administered, followed by the experiment trials. A self-paced rest break was given between Block 1 and Block 2.

The procedure was largely identical to that of Stolz et al. (2005). Each trial began with a fixation symbol (+) displayed in the centre of the screen for 500 ms. A prime word in uppercase letters and Courier font was then presented at fixation for 150 ms. Participants were told to read the prime word silently. A blank interval followed the offset of the prime word. The length of this blank interval varied by condition—it was 50 ms in Experiment 1A, 200 ms in Experiment 1B, and 650 ms in Experiment 1C. The target stimulus was then presented in lowercase Courier font, centred one line below the location previously occupied by the prime. The participants' task was to determine whether the target stimulus spelled a word or nonword, and to respond as quickly and accurately as possible. They were instructed to press the "L" key to indicate that the target was a word, or to press the "A" key to indicate that the target was not a word. The target stimulus remained on the screen until participants made a response. Response times (RTs, to the nearest millisecond) and accuracy were recorded.

Results

Response time (RT) analysis was conducted for critical word target trials on which a correct response was given. Response times were first subjected to a recursive trimming procedure in which the criterion cutoff for outlier removal was established independently for each participant in each condition (van Selst & Jolicoeur, 1994). Outlier removal resulted in 3.09% of the data being excluded in Experiment 1A, 3.76% in Experiment 1B, and 2.95% in Experiment 1C. Following the outlier procedure, the remaining data were subjected to a 2 (Priming: Repeated vs. Unrelated) × 2 (Block: Block 1 vs. Block 2) ANOVA to assess the presence of repetition priming effects. Then, mean repetition priming scores in Block 1 were correlated with mean repetition priming scores in Block 2 for each participant in order to measure the reliability of repetition priming scores. Finally, the reliability of repetition priming scores were compared with the reliability of *semantic* priming scores from Stolz et al. (2005).

Experiment 1A (SOA = 200 ms). Mean RTs and percentage errors for each combination of priming and block are shown in Table 1. Responses were faster and more accurate for repeated targets (542 ms, 4.0%) than for unrelated targets (619 ms, 6.6%), $F(1, 47) = 99.9$, $MSE = 2817$, $p < .001$ for RTs; $F(1, 47) = 25.8$, $MSE = 0.002$, $p < .001$ for errors. Also, responses were more accurate for Block 1 (4.3%) than for Block 2 (6.7%), $F(1, 47) = 9.62$, $MSE = 0.003$, $p < .01$. There was also a marginally significant Block × Priming

TABLE 1
Mean RTs (ms) and percentage errors (%) for repetition priming as a function of relatedness proportion (RP), stimulus–onset asynchrony (SOA), block, and priming

	Block 1			*Block 2*		
Condition	*Repeated*	*Unrelated*	*Priming*	*Repeated*	*Unrelated*	*Priming*
RP = .25						
SOA = 200 ms	551 (3.4)	621 (5.2)	70	534 (4.6)	617 (8.8)	83
SOA = 350 ms	551 (4.7)	631 (10.1)	80	537 (6.2)	602 (11.8)	65
SOA = 800 ms	546 (3.7)	596 (5.1)	50	524 (3.3)	572 (6.3)	48
RP = .50						
SOA = 200 ms	569 (3.1)	674 (7.5)	105	539 (3.1)	665 (7.0)	126
SOA = 350 ms	530 (3.4)	633 (6.8)	103	509 (2.2)	625 (6.4)	116
SOA = 800 ms	562 (4.0)	650 (6.2)	88	525 (3.8)	636 (6.9)	111
RP = .75						
SOA = 200 ms	531 (3.8)	624 (10.4)	93	494 (5.8)	603 (10.6)	109
SOA = 350 ms	516 (5.3)	610 (10.9)	94	490 (5.6)	588 (13.5)	98
SOA = 800 ms	558 (2.6)	634 (6.2)	76	551 (3.6)	613 (9.5)	62

interaction in errors, $F(1, 47) = 3.4$, $MSE = 0.002$, $p = .071$, such that priming effects for Block 1 (1.7%) were marginally smaller than priming effects for Block 2 (4.1%), $t = 1.75$, $p < .10$.

Reliability analysis. Following Stolz et al. (2005), a systematic analysis of the influence of each participant's scores to the overall reliability was conducted for each experiment using studentized residual values. Mean priming scores in Block 2 were regressed onto mean priming scores in Block 1, and data points with studentized residual values larger than 2.5 or smaller than –2.5 were eliminated. In addition, the data points of several participants with an extremely large or extremely small priming score (more than 2.5 standard deviations above or below the mean) were eliminated. The data points that were eliminated are circled in Figure 1.

Following outlier removal, we correlated mean priming scores for each participant in Block 1 with mean priming scores for each participant in Block 2 to determine the reliability of repetition priming effects across

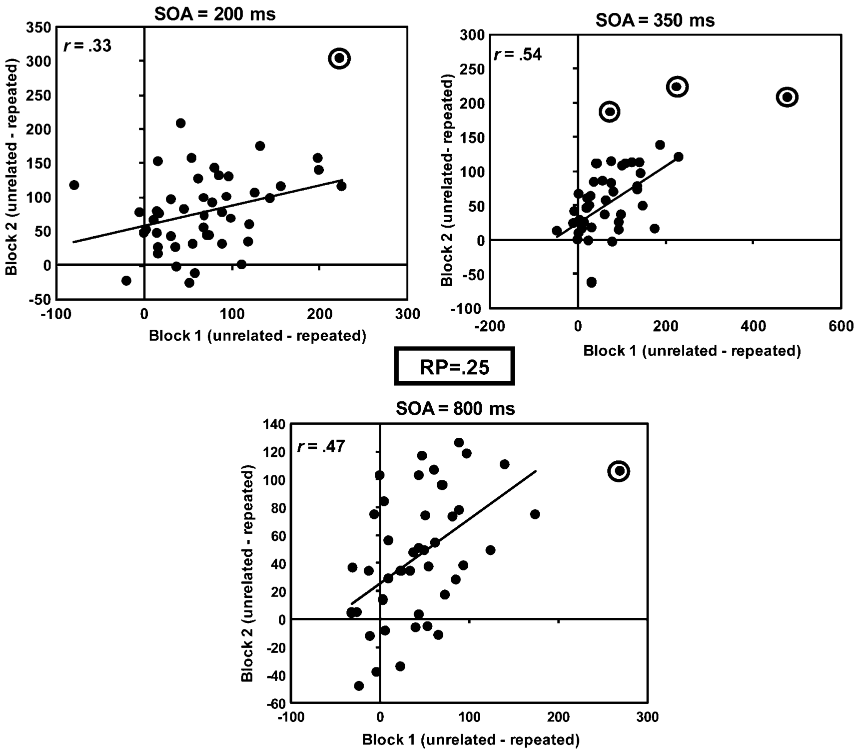

Figure 1. Scatterpoints relating repetition priming scores (RTs for unrelated–repeated critical trials) in Block 1 and Block 2 for Experiments 1A, 1B, and 1C. Outlier values are circled.

blocks. Priming in Experiment 1A showed reliability across blocks, $r = .33$, $t(45) = 5.12$, $p < .05$, $f^2 = .12$ (see Figure 1).

Experiment 1B (SOA = 350 ms). As can be seen in Table 1, responses were faster and more accurate for repeated targets (544 ms, 5.4%) than for unrelated targets (616 ms, 11.0%), $F(1, 47) = 50.9$, $MSE = 4519$, $p < .001$ for RTs; $F(1, 47) = 9.22$, $MSE = 0.016$, $p < .01$ for errors. Overall responses were faster but marginally less accurate in Block 2 (570 ms, 9.0%) than in Block 1 (591 ms, 7.4%), $F(1, 47) = 7.4$, $MSE = 2695$, $p < .01$ for RTs; $F(1, 47) = 3.19$, $MSE = 0.004$, $p = .081$ for errors. There was no Priming × Block interaction, $F = 2.0$, $p = .16$.

Data points were again eliminated for extreme influence and extreme priming scores using the same criteria as in Experiment 1A, and these points appear circled in Figure 1. We again correlated mean priming scores for each participant in Block 1 with mean priming scores for each participant in Block 2. Priming in Experiment 1B showed reliability across blocks, $r = .54$, $t(43) = 4.2$, $p < .001$, $f^2 = .42$ (see Figure 1).

Experiment 1C (SOA = 800 ms). As can be seen in Table 1, responses were faster and more accurate for repeated targets (535 ms, 3.5%) than for unrelated targets (584 ms, 5.7%), $F(1, 47) = 53.6$, $MSE = 2090$, $p < .001$ for RTs; $F(1, 47) = 9.7$, $MSE = 0.002$, $p < .01$ for errors. Overall responses were faster in Block 2 (548 ms) than Block 1 (570 ms), $F(1, 47) = 11.1$, $MSE = 2185$, $p < .01$. There was no Priming × Block interaction ($F < 1$).

Data points were again eliminated for extreme influence and extreme priming scores using the same criteria as in Experiment 1A and 1B, and these points are circled in Figure 1. Mean priming scores for each participant in Block 1 were correlated with mean priming scores for each participant in Block 2. Priming in Experiment 1C showed reliability across blocks, $r = .47$, $t(45) = 3.40$, $p < .01$, $f^2 = .28$ (see Figure 1).

Comparison with semantic priming. In order to determine whether the reliability of repetition priming scores was significantly different from the reliability of *semantic* priming scores (from Stolz et al., 2005; see Table 2), we conducted a null hypothesis test for the difference between correlations in each condition using Fisher's z' transformation.

In Experiment 1A (RP = .25, SOA = 200 ms), we found a correlation of $r = .33$ between scores in blocks, which is significantly different from $r = -.15$ (Stolz et al., 2005) for semantic priming, $z = 2.3$, $p < .05$. In Experiment 1B (RP = .25, SOA = 350 ms), we found a correlation of $r = .54$ between scores across blocks, which is significantly different from $r = .10$ (Stolz et al., 2005) for semantic priming, $z = 2.4$, $p < .01$. In Experiment 1C (RP = .25, SOA = 800 ms), we found a correlation of $r = .47$ between scores in blocks,

TABLE 2
Reliability scores *(r)* for priming across blocks as a function of relatedness proportion (RP) and stimulus–onset asynchrony (SOA)

Condition	*Repetition priming*	*Semantic priming*
RP = .25		
SOA = 200 ms	.33*	−.15
SOA = 350 ms	.54*	.10
SOA = 800 ms	.47*	.08
RP = .50		
SOA = 200 ms	.38*	.38*
SOA = 350 ms	.42*	.28 ($p = .06$)
SOA = 800 ms	−.01	.31*
RP = .75		
SOA = 200 ms	.18	.09
SOA = 350 ms	.43*	.08
SOA = 800 ms	.28 ($p = .06$)	.22*

*$p < .05$.

which is significantly different from $r = .08$ (Stolz et al., 2005) for semantic priming, $z = 2.0$, $p < .05$.

Discussion

The results of Experiment 1 are clear: When RP is low, repetition priming scores are reliable at short, medium, and long SOAs. Repetition priming is also significantly *more reliable than semantic priming* at each of these SOAs.

EXPERIMENT 2

Experiment 2 examined the reliability of repetition priming effects when RP = .50 across the three different SOAs.

Method

Participants. One hundred and forty-four undergraduate students from the University of Waterloo participated in this experiment in exchange for course credit. All participants spoke English as their first language and had normal or corrected-to-normal vision, and none had participated in Experiment 1.

Design. The design was identical to that of Experiment 1.

Stimuli. The stimuli were largely identical to those in Experiment 1, except that 25 repeated and 25 unrelated word-prime word-target buffer pairs were present in each block. Each block therefore contained 50 unrelated word-prime word-target pairs, 50 repeated word-prime word-target pairs, and 100 word prime-nonword target pairs, resulting in an RP of .50.

Procedure. The procedure was identical to that of Experiment 1.

Results

Response time (RT) analysis was conducted for critical word target trials on which a correct response was given. Response times were subjected to the same outlier procedure as in Experiment 1. Outlier removal resulted in 3.33% of the data being excluded in Experiment 2A, 3.06% in Experiment 2B, and 2.91% in Experiment 2C. The same data analysis procedures used in Experiment 1 were used here.

Experiment 2A (SOA = 200 ms). Mean RTs and percent errors for each combination of priming and block are shown in Table 1. Responses were faster and more accurate for repeated targets (554 ms, 3.1%) than for unrelated targets (669 ms, 7.2%), $F(1, 47) = 192.1$, $MSE = 3402$, $p < .001$ for RTs; $F(1, 47) = 35.5$, $MSE = 0.02$, $p < .01$ for errors. Also, responses were faster for Block 1 (602 ms) than for Block 2 (621 ms), $F(1, 47) = 7.22$, $MSE = 246.2$, $p < .05$. There was a significant Block × Priming interaction in RTs, $F(1, 47) = 4.7$, $MSE = 1045$, $p < .05$, such that priming effects for Block 2 (126 ms) were significantly larger than priming effects for Block 1 (105 ms), $t = 2.02$, $p < .05$.

Data points were eliminated for extreme influence and extreme priming scores using the same criteria as in Experiment 1, and these points are circled in Figure 2. To determine the reliability of repetition priming effects across blocks, we correlated mean priming scores for each participant in Block 1 with mean priming scores for each participant in Block 2. Priming in Experiment 2A showed reliability across blocks, $r = .38$, $t(43) = 2.6$, $p < .01$, $f^2 = .17$.

Experiment 2B (SOA = 350 ms). As can be seen in Table 1, responses were faster and more accurate for repeated targets (519 ms, 2.8%) than for unrelated targets (629 ms, 6.6%), $F(1, 47) = 163.7$, $MSE = 3542$, $p < .001$ for RTs; $F(1, 47) = 45.9$, $MSE = 0.02$, $p < .01$ for errors. Overall responses were faster in Block 2 (567 ms) compared with Block 1 (581 ms), $F(1, 47) = 6.04$, $MSE = 1746$, $p < .05$. There was no Priming × Block interaction ($F < 1.2$).

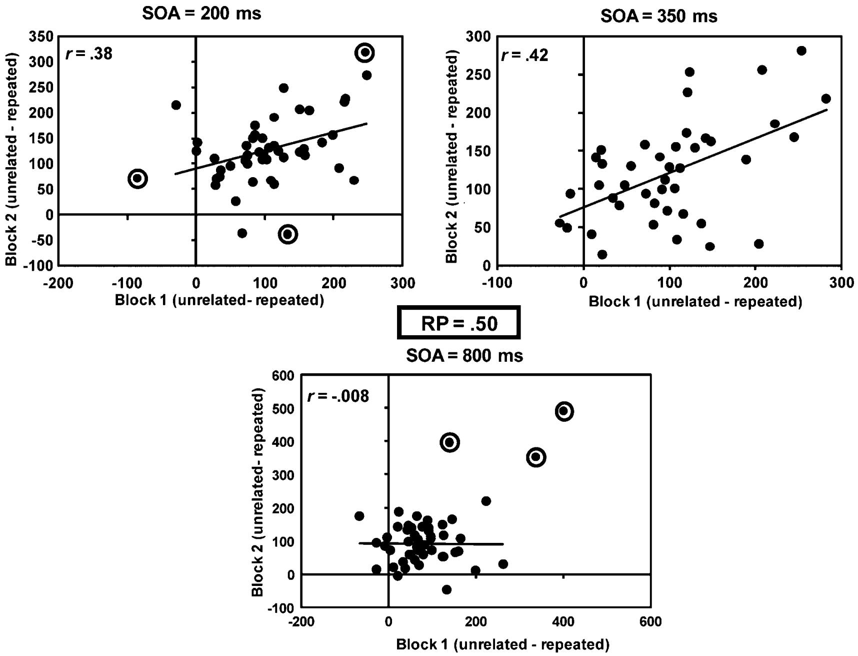

Figure 2. Scatterpoints relating repetition priming scores (RTs for unrelated–repeated critical trials) in Block 1 and Block 2 for Experiments 2A, 2B, and 2C. Outlier values are circled.

Data points were eliminated for extreme influence and extreme priming scores using the same criteria as in Experiment 1, and these points are circled in Figure 2. Mean priming scores for each participant in Block 1 were correlated with mean priming scores for each participant in Block 2. Priming in Experiment 2B showed reliability across blocks, $r = .42$, $t(43) = 3.1$, $p < .01$, $f^2 = .21$ (see Figure 2).

Experiment 2C (SOA = 800 ms). As can be seen in Table 1, responses were faster and more accurate for repeated targets (543 ms, 3.5%) than for unrelated targets (643 ms, 5.7%), $F(1, 47) = 73.5$, $MSE = 6493$, $p < .001$ for RTs; $F(1, 47) = 15.1$, $MSE = 0.002$, $p < .01$ for errors. Overall responses were faster in Block 2 (581 ms) than Block 1 (606 ms), $F(1, 47) = 8.16$, $MSE = 3725$, $p < .01$. There was a marginally significant Block $\times$ Priming interaction in RTs, $F(1, 47) = 2.92$, $MSE = 2119$, $p = .094$, such that the priming effects were marginally larger for Block 2 (111 ms) than for Block 1 (88 ms), $t = 1.71$, $p < .10$.

Data points were again eliminated for extreme influence and extreme priming scores using the same criteria as in Experiment 1, and these points are circled in Figure 2. We again correlated mean priming scores for each participant in Block 1 with mean priming scores for each participant in Block 2. The correlation was not significant, $r = -.008$, $t(43) = 0.055$ (see Figure 2).

Comparison with semantic priming. In order to determine whether the reliability of repetition priming scores was significantly different from the reliability of *semantic* priming scores (from Stolz et al., 2005; see Table 2), we conducted the same null hypothesis tests as in Experiments 1A–1C.

Experiment 2A (RP = .50, SOA = 200 ms) yielded a correlation of $r = .38$ between scores in blocks, which is not significantly different from $r = .38$ for semantic priming, $z = 0$. In Experiment 2B (RP = .50, SOA = 350 ms), there was a correlation of $r = .42$ between scores across blocks, which is not significantly different from $r = .28$ for semantic priming, $z = .76$. In Experiment 2C (RP = .50, SOA = 800 ms), there was a correlation of $r = -.008$ between scores in blocks, which was marginally smaller than $r = .31$ for semantic priming, $z = 1.6$.

Discussion

These data indicate that repetition priming yielded significant reliability at an RP of .50 with short and medium SOAs. When the SOA is long (800 ms), however, individuals' repetition priming scores for Block 1 yielded effectively zero correlation with their repetition priming scores for Block 2. Possible explanations for the observed lack of reliability at the long SOA are considered in the General Discussion.

EXPERIMENT 3

Experiment 3 examines the reliability of repetition priming effects when RP = .75 across three SOAs.

Method

Participants. One hundred and forty-four undergraduate students from the University of Waterloo participated in this experiment in exchange for course credit. All participants spoke English as their first language and had normal or corrected-to-normal vision, and none had participated in Experiment 1 or 2.

Design. The design was identical to that of Experiment 1 and 2.

Stimuli. The stimuli were largely identical to those in Experiment 1 and 2, except that 50 repeated word-prime word-target buffer pairs were present in each block. Each block then contained 25 unrelated word-prime word-target pairs, 75 repeated word-prime word-target pairs, and 100 word prime-nonword target pairs, resulting in an RP of .75.

Procedure. The procedure was identical to that of Experiment 1 and 2.

Results

Response time (RT) analysis was conducted for critical word target trials on which a correct response was given. Response times were subjected to the same outlier procedure as in Experiments 1 and 2. Outlier removal resulted in 3.80% of the data being excluded in Experiment 2A, 3.27% in Experiment 2B, and 3.67% in Experiment 2C. The same data analysis procedures used in Experiments 1 and 2 were used here.

Experiment 3A (SOA = 200 ms). Mean RTs and percent errors for each combination of priming and block are shown in Table 1. As can be seen in the table, responses were faster and more accurate for repeated targets (512 ms, 5.8%) than for unrelated targets (613 ms, 10.5%), $F(1, 47) = 144.0$, $MSE = 3380$, $p < .001$ for RTs; $F(1, 47) = 26.5$, $MSE = 0.006$, $p < .01$ for errors. Also, responses were faster for Block 1 (548 ms) than for Block 2 (577 ms), $F(1, 47) = 9.26$, $MSE = 4318$, $p < .05$. There was no Priming × Block interaction ($F < 1$).

Data points were eliminated for extreme influence and extreme priming scores using the same criteria as in Experiments 1 and 2, and these points are circled in Figure 3. To determine the reliability of repetition priming effects across blocks, mean priming scores for each participant in Block 1 were correlated with mean priming scores for each participant in Block 2. The correlation was not significant, $r = .18$, $t(44) = 1.2$, $p > .2$, $f^2 = .03$.

Experiment 3B (SOA = 350 ms). As can be seen in Table 1, responses were faster and more accurate for repeated targets (503 ms, 5.4%) than for unrelated targets (599 ms, 12.2%), $F(1, 47) = 126.3$, $MSE = 3527$, $p < .001$ for RTs; $F(1, 47) = 48.7$, $MSE = 0.004$, $p < .01$ for errors. Overall responses were faster in Block 2 (539 ms) compared with Block 1 (563 ms), $F(1, 47) = 7.81$, $MSE = 3407$, $p < .05$. There was no Priming × Block interaction (Fs < 1).

Data points were eliminated for extreme influence and extreme priming scores using the same criteria as in Experiment 1, and these points appear

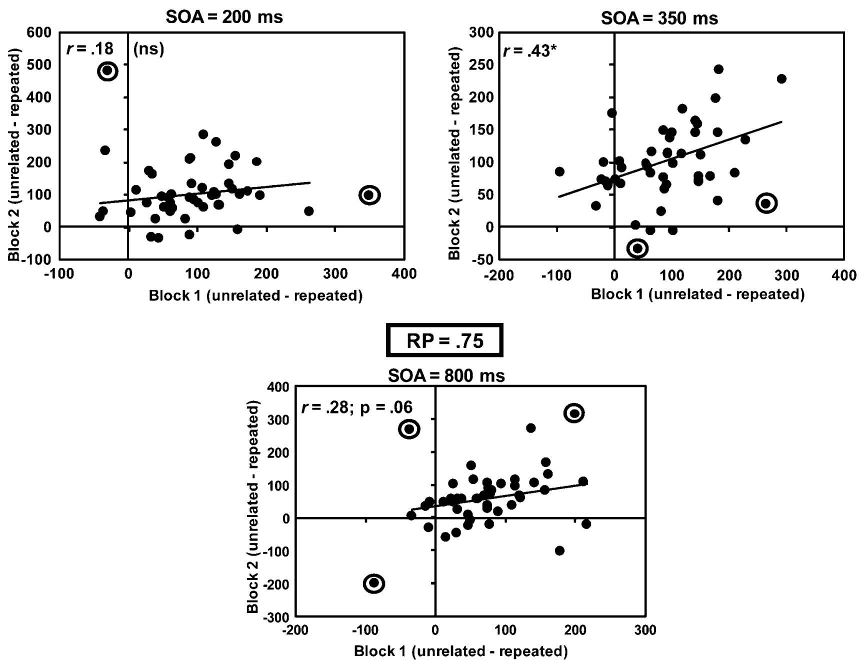

Figure 3. Scatterpoints relating repetition priming scores (RTs for unrelated–repeated critical trials) in Block 1 and Block 2 for Experiments 3A, 3B, and 3C. Outlier values are circled.

circled in Figure 3. Mean priming scores for each participant in Block 1 were correlated with mean priming scores for each participant in Block 2. Priming in Experiment 2B showed reliability across blocks, $r = .43$, $t(44) = 6.4$, $p < .001$, $f^2 = .22$ (see Figure 3).

Experiment 3C (SOA = 800 ms). As can be seen in Table 1, responses were faster and more accurate for repeated targets (555 ms, 3.1%) than for unrelated targets (623 ms, 7.9%), $F(1, 47) = 54.5$, $MSE = 4122$, $p < .001$ for RTs; $F(1, 47) = 31.5$, $MSE = 0.004$, $p < .01$ for errors. Overall responses were marginally faster and significantly *less* accurate in Block 2 (582 ms, 6.5%) than Block 1 (596 ms, 4.4%), $F(1, 47) = 3.96$, $MSE = 2428$, $p = .052$ for RTs; $F(1, 47) = 13.9$, $MSE = 0.002$, $p = .001$. There was no Priming × Block interaction ($F < 1$).

Data points were eliminated for extreme influence and extreme priming scores using the same criteria as in Experiment 1, and these points appear circled in Figure 3. Mean priming scores for each participant in Block 1 were correlated with mean priming scores for each participant in Block 2.

Priming in Experiment 3C showed marginally significant reliability across blocks, $r = .28$, t $(43) = 1.90$, $p = .06$, $f^2 = .01$.

Comparison with semantic priming. We conducted the same null hypothesis tests as in Experiments 1 and 2. In Experiment 3A (RP = .75, SOA = 200 ms), we found a correlation of $r = .18$ between scores in blocks, which is not significantly different from $r = .09$ for semantic priming, $z = .43$. In Experiment 3B (RP = .75, SOA = 350 ms), we found a correlation of $r = .43$ between scores in blocks, which was significantly higher than the $r = .08$ for semantic priming, $z = 1.8$, $p < .05$. In Experiment 3C (RP = .75, SOA = 800 ms), we found a correlation of $r = .28$ between scores in blocks, which is not significantly different from $r = .22$ for semantic priming, $z = .32$.

Discussion

The data from Experiment 3, in which RP = .75, indicate that repetition priming exhibits significant reliability at medium (350 ms) SOAs and long (800 ms) SOAs, but not at short (200 ms) SOAs.

GENERAL DISCUSSION

First, the present experiments serve to highlight the impressive general robustness of repetition priming, which was present at the group level in every RP and SOA condition. However, it is interesting to note that even repetition priming was not universal on an individual level. Of the 432 participants in this series of experiments, 27 showed a net priming effect equal to or less than zero. That is, roughly 6% of our participants showed no net repetition priming effects at all. Although this is appreciably lower than the 15% of participants who showed no net *semantic* priming effects in the Stolz et al. (2005) study, it still serves to emphasize the dramatic variance in the size of our effects that occurs at the level of the individual. This variance, too often ignored in cognitive psychology, may represent a rich ground for experimentation and theory building.

We turn now to a discussion of the two main objectives of this series of experiments. The first purpose was to compare the reliability of repetition priming with the reliability of semantic priming, in order to determine whether the observed unreliability of semantic priming is specific to *semantic*-level processes. The second purpose was to examine the pattern of reliability across the different RP and SOA conditions, and to establish what this pattern may tell us about the reliability of automatic and controlled processes in visual word recognition.

Comparison with semantic priming

With the exception of one condition, repetition priming scores across blocks were just as reliable (four conditions) or significantly more reliable (four conditions) than semantic priming scores across blocks. With one exception, there were no conditions in which semantic priming was found to be significantly more reliable than repetition priming. However, when RP was .50 and the SOA was long, the reliability of semantic priming was marginally more reliable then the reliability of repetition priming.

One might wonder whether the higher reliability of repetition priming (as compared to semantic priming) reflects a restriction of range problem.[1] Computing correlations depends strongly on the amount of variance in the sample, and low variance in the priming scores might result in difficulties detecting correlations that would be present if variance were greater. We therefore compared the standard deviations and ranges of repetition priming scores with the standard deviations and ranges of semantic priming scores from Stolz et al. (2005). These data are presented in Table 3. The ranges across the two paradigms are relatively comparable; however, the standard deviations are slightly but consistently larger for repetition priming than for semantic priming. It is important to note that the variability in semantic priming scores is still robust. Furthermore, the differences in *variability* do not seem to map onto the differences in *reliability* in any consistent way; some of the largest discrepancies in variability between paradigms occur in conditions where no differences in reliability were found (for example, in the RP = .50, SOA = 350 ms condition).

In general, therefore, repetition priming is more reliable than semantic priming. We can consequently conclude, as Stolz et al. (2005) suggested, that the observed unreliability of semantic priming results largely from uncoordinated processes specific to semantic memory. Lexical and prelexical processes in visual word recognition are more reliable and unfold in a more coordinated, consistent manner.

The pattern of reliability

How do we account for the *absence* of reliability in two of the nine experimental conditions? We consider several possibilities: The low reliability of difference scores, a potential restriction of range problem, and the interplay of automatic and controlled processes.

[1] Charles Folk and an anonymous reviewer brought this point to our attention.

TABLE 3
Standard deviations and ranges for semantic priming and repetition priming by condition after removing outliers

	Repetition priming			*Semantic priming (Stolz et al., 2005)*		
Condition	*SD of priming*	*Min to max priming*	*Priming range*	*SD of priming*	*Min to max priming*	*Priming range*
RP = 25						
SOA = 200	46	−21 to +177	198	35	−43 to +161	204
SOA = 350	48	−17 to +175	192	36	−129 to +120	249
SOA = 800	39	−36 to +125	161	30	−58 to +101	159
RP = .50						
SOA = 200	51	+15 to +261	246	42	−88 to +123	211
SOA = 350	60	+14 to +268	254	37	−71 to +146	217
SOA = 800	43	−8 to +220	228	27	−16 to +107	123
RP = .75						
SOA = 200	53	−6 to +198	204	38	−53 to +124	177
SOA = 350	58	−5 to +198	203	39	−37 to +163	200
SOA = 800	50	−29 to +204	233	41	−44 to +206	250

Low reliability of difference scores.. In order to compute reliability, we correlated participants' repetition priming scores for Block 1 (RTs for unrelated targets–RTs for repeated targets) with their repetition priming scores for Block 2 (RTs for unrelated targets–RTs for repeated targets). Computing reliability, therefore, involves correlating *difference scores*, which can be problematic (Williams & Zimmerman, 1996). The reliability of difference scores is limited by the reliability of the scores used to compute them, and is therefore inherently less reliable than other types of measures. It is therefore possible that the observed unreliability of repetition priming scores arises from the methods used to compute reliability and simply reflects an inability to detect reliability that is actually present. However, the reliability of priming scores was computed in exactly the same way for all nine of our experimental conditions, and problems with the reliability of difference scores cannot easily account for the observed *differences* in reliability across conditions.

Restriction of range. Perhaps the low reliability of priming in some conditions simply reflects a restriction of range problem. We therefore closely examined the standard deviations and ranges in the two conditions where reliability was absent (see Table 3). The ranges and standard deviations of the priming scores for the two conditions where reliability was *not* observed were similar to the ranges and standard deviations in other experimental conditions where reliability *was* observed. Therefore, the differences in reliability across condition cannot simply be explained by a restriction of range problem.

Automatic and controlled processing. There is a large literature on the mechanisms that produce priming, and various processes have been posited. These processes are summarized in Stolz et al. (2005). Automatic processes are the chief contributors to priming at low RPs and short SOAs, when there is neither time nor incentive to recruit more resource-dependent and strategic processes. Controlled and effortful processes, however, are sensitive to participants' goals and knowledge, and require both time and attentional resources to implement (Stolz et al., 2005). Therefore, controlled and effortful processes are most likely to be recruited at high RPs and long SOAs.

There is some evidence that the interplay of controlled and automatic processes influence reliability. For example, Borgmann, Risko, Stolz, and Besner (2007) examined the reliability of the Simon effect (Simon, 1990). The Simon effect refers to the finding that participants identify targets more quickly when the irrelevant spatial location of the target (e.g., left) is compatible with the location of the response key (e.g., left) compared with when the location of the response key is incompatible (e.g., right). Borgmann et al. found that the reliability of the Simon effect increased monotonically with the proportion of compatible trials. Because changes in the proportions of compatible/congruent trials lead to changes in the recruitment of controlled and strategic processes (see Kane & Engle, 2003), Borgmann et al. attributed differences in reliability across conditions to the recruitment of more controlled processes as the proportion of compatible trials increased.

Also, Stolz et al. (2005) explained the observed differences in reliability in semantic priming scores across experimental conditions in terms of differential contribution of controlled and automatic processes. With increasing RP, the retrieval of prime episodes (e.g., Bodner & Masson, 2001) became increasingly systematic, resulting in increased reliability. When RP increased to .75, however, participants began to engage in the controlled, strategic generation of potential target candidates (e.g., Neely, 1977; Posner & Snyder, 1975). When the SOAs were short, these expectancies did not have time to succeed, resulting in low reliability.

Though the processes that contribute to *repetition* priming at different RPs and SOAs are less well-defined, some of the mechanisms may be similar. For example, Bodner and Masson (2001) present an episodic resource account of masked priming that may also apply to both unmasked repetition priming and to semantic priming. Under this account, the prime episode is encoded automatically and is recruited retrospectively in a manner contingent on how useful the prime episode is for target processing—that is, when RP is low, the prime episode is recruited less often or less consistently than when RP is high. The prime episode may also be recruited less often or less consistently in the semantic priming paradigm than in the repetition priming paradigm (because the prime episode is more useful in the repetition priming paradigm). Therefore, perhaps the higher general reliability of repetition priming compared to semantic priming results from more consistent use of episodic retrieval processes.

Similarly, a strategic backward matching process (e.g., Neely, Keefe, & Ross, 1989) might require less effort in a repetition priming paradigm than a semantic priming paradigm (i.e., checking whether "BOAT" is the same as "boat" may require less effort than checking whether "ship" is semantically related to "boat"). As such, participants may engage in strategic backward matching more consistently in a repetition priming paradigm, leading to higher reliability for repetition priming compared to semantic priming.

The interplay of controlled and automatic processes might also explain the absence of reliability in two of our experimental conditions. For example, the absence of reliability when RP was high (RP = .75) and SOA was short (SOA = 200 ms) may reflect participants' attempts to use controlled and effortful processes. At the shortest SOA, these processes may not have had time to succeed, resulting in the observed low reliability of priming scores in that condition.

Implications for individual and group differences research

The present results have important implications for researchers who use repetition priming to study individual and group differences (for example, in implicit memory). Though repetition priming did show significant reliability across blocks in most conditions, this reliability was generally relatively modest (ranging from 0 to .54). Therefore, though repetition priming scores under most conditions *do* exhibit systematic variance that can and should be explained, the relatively low reliability of this measure can attenuate the power to detect correlations and between-group differences. Researchers should therefore be cautious in interpreting any null effects, and should bear in mind that they may need large groups of participants to detect any meaningful differences between groups.

SUMMARY AND CONCLUSION

The present study consisted of a systematic examination of the reliability of individual participants' repetition priming across blocks. The results indicate that repetition priming is reliable under most RP and SOA conditions, with two exceptions: When RP is .50 and SOA is long (800 ms), and when RP is .75 and SOA is short (200 ms), repetition priming is not reliable. These results have two major implications. First, repetition priming is largely more reliable than semantic priming, suggesting that the observed unreliability of semantic priming arises from uncoordinated processes specific to semantic memory. Second, the interplay of automatic and controlled processes likely contributes to the differences in reliability across conditions. Researchers who use repetition priming as a tool to study individual and between-group differences should keep reliability in mind and choose conditions and sample sizes appropriately.

REFERENCES

Balota, D. A., Black, S. R., & Cheney, M. (1992). Automatic and attentional priming in young and older adults: Re-evaluation of the two-process model. *Journal of Experimental Psychology: Human Perception and Performance*, *18*, 485–502.

Balota, D. A., & Duchek, J. M. (1991). Semantic priming effects, lexical repetition effects, and contextual disambiguation effects in healthy aged individuals and individuals with senile dementia of the Alzheimer's type. *Brain and Language*, *40*, 181–201.

Bodner, G. E., & Masson, M. E. J. (2001). Prime validity affects masked repetition priming: Evidence for an episodic resource account of priming. *Journal of Memory and Language*, *45*, 616–647.

Borgmann, K. W. U., Risko, E. F., Stolz, J. A., & Besner, D. (2007). Simon says: Reliability and the role of working memory and attentional control in the Simon task. *Psychonomic Bulletin and Review*, *14*, 313–319.

Bowers, J. S. (2000). In defense of abstractionist theories of word identification and repetition priming. *Psychonomic Bulletin and Review*, *7*, 83–99.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*, 204–256.

Danion, J. M., Willard-Schroeder, D., Zimmerman, M. A., Grange, D., Schlienger, J. L., & Singer, L. (1991). Explicit memory and repetition priming in depression. *Archives of General Psychiatry*, *48*, 707–711.

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

Jacoby, L. L. (1983). Remembering the data: Analyzing interactive processes in reading. *Journal of Verbal Learning and Verbal Behavior*, *22*, 485–508.

Kane, M. J., & Engle, R. W. (2003). Working-memory capacity and the control of attention: The contributions of goal neglect, response competition, and task set to Stroop interference. *Journal of Experimental Psychology: General*, *132*, 47–70.

Kopriva, R. J., & Shaw, D. G. (1991). Power estimates: The effect of dependent variable reliability on the power of one-factor ANOVAs. *Educational and Psychological Measurement*, *51*, 585–595.

Madden, D. J., Pierce, T. W., & Allen, P. A. (1993). Age-related slowing and the time-course of semantic priming in visual word identification. *Psychology and Aging*, *8*, 490–507.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: An account of basic findings. *Psychological Review*, *88*, 375–407.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, *76*, 165–178.

Nebes, R. N., Brady, C. B., & Huff, F. J. (1989). Automatic and attentional mechanisms of semantic priming in Alzheimer's disease. *Journal of Clinical and Experimental Neuropsychology*, *11*, 219–230.

Neely, J. H. (1997). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General, 106*, 226–254.

Neely, J. H., Keefe, D. E., & Ross, K. (1989). Semantic priming in the lexical decision task: Roles of prospective prime-generated expectancies and retrospective semantic matching. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 1003–1019.

Perry, C., Zeigler, J. C., & Zorzi, M. (2007). Nested incremental modeling in the development of computational theories: The CDP+ model of reading aloud. *Psychological Review*, *114*, 273–315.

Posner, M. I., & Snyder, C. R. R. (1975). Facilitation and inhibition in the processing of signals. In P. M. A. Rabbitt & S. Dornic (Eds.), *Attention and performance V* (pp. 669–682). New York: Academic Press.

Scarborough, D. L., Cortese, C., & Scarborough, H. F. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance*, *3*, 1–17.

Simon, J. R. (1990). The effects of an irrelevant directional cue on human information processing. In R. W. Proctor & T. G. Reeve (Eds.), *Stimulus–response compatibility: An integrated perspective* (pp. 31–86). Amsterdam: North-Holland.

Stolz, J. A., & Besner, D. (1996). Role of set in visual word recognition: Activation and activation blocking as a nonautomatic process. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 1166–1177.

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.

Tenpenny, P. L. (1995). Abstractionist versus episodic theories of repetition priming and word identification. *Psychonomic Bulletin and Review*, *2*, 339–363.

Van Selst, M., & Jolicoeur, P. (1994). A solution to the effect of sample size on outlier elimination. *Quarterly Journal of Experimental Psychology*, *47A*, 631–650.

Williams, R. H., & Zimmerman, D. W. (1996). Are simple gain scores obsolete? *Applied Psychological Measurement*, *20*, 59–69.

Manuscript received October 2008
Manuscript accepted March 2009
First published online June 2009