Running head: STUDY DURATION, RECOGNITION, AND PRIMING

Does Study Duration Have Opposite Effects on Recognition and Repetition Priming?

Christopher J. Berry

Plymouth University

Emma V. Ward

Middlesex University

David R. Shanks

University College London

Author Note

Christopher J. Berry, School of Psychology, Plymouth University; Emma V. Ward, Psychology Department, Middlesex University; David R. Shanks, Division of Psychology and Language Sciences, University College London.

Correspondence concerning this article should be addressed to Christopher Berry, School of Psychology, Plymouth University, Drake Circus, PL4 8AA, UK. christopher.berry@plymouth.ac.uk.

Abstract

We investigated whether manipulating the duration for which an item is studied has opposite effects on recognition memory and repetition priming, as has been reported by Voss and Gonsalves (2010). Robust evidence of this would support the idea that distinct explicit and implicit memory systems drive recognition and priming, and would constitute evidence against a single-system model (Berry, Shanks, Speekenbrink, & Henson, 2012). Across seven experiments using study durations ranging from 40 ms to 2250 ms, and two different priming tasks (a classification task in Experiments 1a, 2a, 3a, and 4, and a continuous identification with recognition (CID-R) task in Experiments 1b, 2b, and 3b), we found that although a longer study duration improved subsequent recognition in each experiment, there was either no detectable effect on priming (Experiments 1a, 2a, and 4) or a similar effect to that on recognition, albeit smaller in magnitude (Experiments 1b, 2b, 3a, and 3b). Our findings 1) question whether study duration has opposite effects on recognition and priming, and 2) are robustly consistent with a single-system model of recognition and priming.

Keywords: recognition memory; repetition priming; study duration; mathematical model; signal detection theory.

Comparisons of recognition memory and long-term repetition priming have played a major role in the development of theories of the organization of memory (Tulving & Schacter, 1990; Squire & Dede, 2015). Recognition memory refers to the capacity to judge whether an item (e.g., a word or object) has been presented before in a particular context. Long-term repetition priming (henceforth priming) refers to a change in identification, detection, or production of an item, which occurs as a result of prior exposure to the same or a similar item. This change is often evident as an improvement in performance and can persist over minutes or longer (and so can be considered *long-term*). For example, identification latencies of objects that have been presented before in a study phase tend to be shorter than those of novel nonpresented items. Individuals with amnesia, arising from damage to the medial temporal lobes/hippocampus, show marked deficits in recognition memory, and yet their capacity to show priming can be left relatively intact, compared to normal adults (e.g., Hamann & Squire, 1997; Schacter, Chiu, Ochsner, 1993). Various experimental manipulations have also been shown to differentially affect recognition and priming in healthy individuals and, together with the findings from amnesic individuals, have been used to support the now widely held *multiple-systems* view that recognition and priming are driven by functionally and neurally distinct explicit and implicit memory systems in the brain (Gabrieli, 1998; Squire, 2004; Squire & Dede, 2015; Tulving & Schacter, 1990).

Despite decades of research, the idea that there exists a sharp distinction between explicit and implicit memory systems is still disputed (see e.g., Addante, 2015; Berry, Shanks, Speekenbrink, & Henson, 2012; Dew & Cabeza, 2011; Hannula & Greene, 2012; Henke, 2010; Reder, Park, & Kieffaber, 2009; Shanks & Berry, 2012). Recognition memory and priming, in particular, may not be independent from one another as once thought. Many experimental factors that were initially believed to selectively affect either recognition or priming (providing evidence for a single dissociation) have since been shown to have similar

effects on recognition and priming. This has been shown, for example, with the effects of normal aging (Ward, Berry, & Shanks, 2013a, 2013b), divisions of attention at encoding (Berry, Henson, & Shanks, 2006), retroactive interference (Eakin & Smith, 2012), changes in presentation modality between study and test phases (Craik, Moscovitch, & McDowd, 1994; Mulligan & Osborn, 2009), levels of processing (Brown & Mitchell, 1994), and also amnesia (Berry et al., 2014; Ostergaard, 1999). This highlights a well-known limitation with the use of single dissociations (and Null Hypothesis Significance Testing) as evidence for multiple systems, which is that they rely on concluding that an effect of a variable on either recognition or priming is absent. Such a conclusion is problematic given that the variable may actually have an effect that, in reality, is relatively small and hard to detect, particularly if the sensitivity of the task is relatively low (Buchner & Wippich, 2000; Dunn, 2003). The same limitation applies when two single dissociations are used together to provide evidence of a double dissociation (see Dunn, 2003).

Stronger support for the notion that recognition and priming are driven by multiple systems would be a crossover dissociation, that is, a demonstration that an independent variable has opposite effects on recognition and priming. Such evidence, however, is rare. One classic example was reported by Jacoby (1983), who found that generating a target word from its antonym in the study phase (i.e., generate the word ‘cold’ from the cue ‘hot-???’) led to greater subsequent recognition of the target compared to when a target word had simply been read during study. In contrast, the same encoding manipulation produced less priming (in a perceptual identification task) for items that were generated rather than read. This pattern has been replicated by others (e.g., Blaxton, 1989; Masson & MacLeod, 1992) and has also been demonstrated with auditory stimuli (e.g., Dew & Mulligan, 2008), and suggests that recognition and priming rely on different sources of information. However, this dissociation can alternatively be interpreted in terms of the principle of transfer appropriate

processing, whereby recognition and priming in the perceptual identification task differentially rely upon conceptual and perceptual processes, rather than distinct explicit and implicit memory systems (Blaxton, 1989; Jacoby, 1983). The idea is that generating a word evokes conceptual processing, which supports greater performance on a recognition task that draws heavily upon this type of processing. Conversely, reading a word evokes perceptual processing, supporting greater perceptual priming for the item. The production of a crossover dissociation using a read-generate manipulation at encoding also seems to critically depend upon words being generated from antonyms at encoding, since other methods of generating targets do not produce a reversal in the generation effect in priming (see Mulligan & Dew, 2009). On the whole, read-generate manipulations only produce a crossover dissociation between recognition and priming under very specific conditions, and, even when produced, may not necessarily reflect the operations of distinct explicit and implicit memory systems.

Other evidence for a crossover dissociation was more recently provided by Voss and Gonsalves (2010), who reported that study duration has opposite effects on recognition and priming. In the study phase of their experiment, participants classified pictures of objects presented for a brief (250 ms) or long duration (2000 ms) as natural or manmade. In the test phase, participants were presented with previously studied pictures of objects (half that were previously presented for a brief duration, and half for a long duration), and new objects for 500 ms, and once again were asked to classify the items as natural or manmade. The priming effect was calculated as the mean classification RT to new items minus the mean classification RT to old items (brief or long). After each classification, an old/new recognition judgment was made. Significant priming effects were found for both brief and long items, but, crucially, the mean priming effect was 19 ms (95% CI [6, 32], Cohen's d_z = 0.855, estimated from the results in Voss and Gonsalves) greater for brief items than for long items. The recognition results showed the opposite pattern: the proportion of long items

correctly judged old (hits) was significantly greater than the proportion of brief items correctly judged old. (Both long and brief items were also judged old more often than new items.)

Voss and Gonsalves (2010) also measured event-related potential (ERP) responses during the test phase. The main findings here were that ERPs to objects that had been studied for a brief duration were more negative than those for new objects at parieto-occipital electrodes in the 200-400 ms interval after the stimulus onset at test, and this was not observed for long items. Instead, ERPs to objects that had been studied for a long duration were more positive in central-parietal electrodes in the 400-600 ms interval, relative to new objects, and this was not found for brief items. Given the behavioural differences between priming and recognition for brief and long conditions, the early negative repetition effects for brief items were attributed to priming, and the later positive effects were attributed to explicit remembering. The overall findings were taken as evidence against a single-system view of recognition and priming, and instead were taken to support a multiple systems view in which study duration differentially engages independent explicit and implicit memory systems at encoding.

Earlier studies looked at the effects of study duration on recognition and priming, but the majority used relatively long durations even in the brief condition (i.e., 1s or longer), and reported single dissociations in which longer durations improved recognition but had little or no effect on priming (e.g., comparing durations of 1 s vs. 3 s in Jacoby & Dallas, 1981; 1s vs. 10 s in Musen, 1991; 1s, 3s, vs. 6.5 s in Neill, Beck, Bottalico, & Molloy, 1990). von Hippel and Hawkins (1994) used study durations shorter than 1s, and found that performance in two explicit memory tasks (graphemic and semantic cued recall) and three implicit memory tasks (word fragment completion, perceptual identification and general knowledge) tended to be better as study duration increased from 50 ms to 2000 ms, and no dissociation was observed

in their study. The authors did not include a recognition task, however, and recognition memory might reasonably be expected to increase across such brief durations. Indeed, Wichmann, Sharpe, and Gegenfurtner (2002) found that recognition memory (for scenes) increased reliably across study durations of 50 to 1067 ms, though this study did not include a priming measure.

Although a few studies have found that priming is not a monotonically increasing function of study duration (e.g., Zago, Fenske, Aminoff, & Bar, 2005; Miyoshi & Ashida, 2014; Miyoshi, Kimura, & Ashida, 2015), these studies either did not additionally examine recognition (Miyoshi et al., 2015; Zago et al., 2005), or, if recognition was also examined, found no evidence of a dissociation (Miyoshi & Ashida, 2014; though this study was particularly focused on recognition accuracy for guesses). These studies are discussed further in the *General Discussion*. The study by Voss and Gonsalves (2010) is, to our knowledge, the first to demonstrate that study duration produces a crossover dissociation between recognition and priming when the encoding conditions are identical for both tasks. We regard such demonstrations as more compelling than comparisons of individual priming or recognition conditions from different studies because they help to limit the range of alternative explanations of the dissociation (see also Ryan & Cohen, 2003).

Why might a brief study duration cause greater priming than a long one? Voss and Gonslaves (2010) offered two potential explanations. A study duration of approximately 250 ms might be optimal if priming is driven by neural "sharpening" and "selection" processes. Indeed Zago et al. (2005) observed a duration-dependent rise and fall of cortical brain deactivation in a functional magnetic resonance imaging (fMRI) study of priming. An alternative transfer-appropriate processing hypothesis is that the rapid perceptual processing required to identify a briefly-presented study item, as compared to a long-duration study item, transfers better to the test phase, which also requires rapid identification.

If study duration has opposite effects on recognition and priming then this would pose a serious challenge for single-system theories of recognition and priming. Indeed, Voss and Gonsalves (2010) concluded that their findings are difficult to reconcile with a single-system view, such as the model proposed by Berry et al. (2008) in which recognition and priming are driven by the same underlying memory system or signal. This model, and reasons why such a crossover dissociation presents a challenge, are outlined in the next section.

The Single-System Model and Predictions

The single-system model has been shown to predict many results concerning recognition and priming, and has even been shown to outperform numerous multiple-systems versions of the model (see Berry et al., 2008; Berry et al., 2012; Berry et al., 2014). The model is outlined here in order to formally describe its relevant predictions and to explain why it does not predict a crossover dissociation. The single-system model assumes that each item at test is associated with a memory strength variable *f*, which is a normally distributed, random variable with mean μ_X and standard deviation σ_f (i.e., $f \sim N(\mu_X, \sigma_f)$), where the subscript *X* denotes the stimulus type (e.g., old, new, brief, long etc.). To generate a response for an item in the recognition task, its value of *f* is first added to a noise variable, e_r, which is also a normally distributed random variable with a mean of zero (i.e., $M(e_r) = 0$) and standard deviation σ_r (i.e., $e_r \sim N(0, \sigma_r)$). Summing *f* with e_r gives J_r:

$$J_r = f + e_r. \qquad [1].$$

An item's value of J_r is then compared to a response criterion, *C* (a scalar), and, as in signal detection models (Macmillan & Creelman, 2005), if J_r exceeds *C*, then the item is judged old, or else it is judged new. Because old items have been presented in the study phase, they will tend to have a greater mean *f* than new items at test. When μ_{old} is greater than μ_{new} (i.e., μ_{old} -

$\mu_{new} > 0$), old items will tend to be judged old more often than new items (i.e., the hit rate will be greater than the false alarm rate), and so a measure of sensitivity (e.g., d') will be greater than the level expected according to chance (e.g., $d' > 0$). The mean f of new items is set to equal zero (i.e., $\mu_{new} = 0$), and hence sensitivity is directly related to the value of μ_{old} (because $\mu_{old} - \mu_{new} = \mu_{old} - 0 = \mu_{old}$). In previous applications of the model (Berry et al., 2012, 2014), the values of σ_f and σ_r were set to equal $\sqrt{0.5}$, and the expected value of d' is therefore equal to μ_{old} (because $d' = (\mu_{old} - \mu_{new}) / \sigma_{Jr}$, and $\sigma_{Jr} = \sqrt{(\sigma_f^2 + \sigma_r^2)} = 1$).

The same value of f that was used to generate an item's response in the recognition task is used to generate its response in the priming task. This assumption is what makes the model a single-system model—that the recognition and priming measures of an item are calculated from the same value of the memory signal f. Crucially, however, f is subjected to another and independent source of noise, e_p, in the generation of a priming task response, where e_p is another normally distributed random variable with a mean of zero (i.e., $M(e_p) = 0$) and standard deviation σ_p (i.e., $e_p \sim N(0, \sigma_p)$). For example, in a task where priming effects are measured using response times (RTs), such as an identification or classification task, f can be combined with e_p to give the RT as follows:

$$RT = b - sf + e_p, \qquad [2]$$

where the parameters b and s are scalars: b is the RT intercept and s is the rate of change in RT with f; the parameter s also serves to scale the variance of e_p (because Equation 2 could be rewritten equivalently as $RT = b - s(f + e_p)$ with e_p suitably rescaled). The influence of f on RT is therefore such that RT is a decreasing function of f in Equation 2—the greater the f value of an item, the faster the RT. Because old items have a greater mean f than new items, the RT for old items will tend to be faster than the RT of new items and the model will

produce a priming effect. It follows from Equation 2 that the expected *RT* of an item is $b - s\mu_X$. For new items, because $\mu_{new} = 0$, the expected RT is $b - s(0) = b$. For old items, the expected *RT* is $b - s\mu_{old}$. The expected priming effect is given as the difference in expected *RT* to new and old items, $b - (b - s\mu_{old}) = s\mu_{old}$. Thus, as with recognition memory, the magnitude of priming is positively related to μ_{old}.

Varying μ_{old} will therefore produce similar effects on both recognition (for which predicted $d' = \mu_{old}$) and priming ($= s\mu_{old}$), but because there are differences in the way that *f* is transformed for each task, the rate of change in priming and recognition with μ_{old} is not necessarily the same (e.g., if standardised effect sizes are computed and compared). The effect on priming could even be so small as to often go empirically undetected (Berry et al., 2006; Dunn, 2003). If μ_{old} were greater for items presented for a longer duration at study compared to a brief duration, the model would predict that both priming and recognition will increase with study duration. Thus, if study duration had opposite effects on recognition and priming, then this would be strong evidence against the model.

From Equations 1 and 2 it follows that J_r and *RT* will be jointly distributed as a bivariate normal distribution with covariance equal to $-s\sigma f^2$ (and mean vector equal to $[\mu_{old}, s\mu_{old}]$, if μ_{new} is fixed to zero). The negative covariance between J_r and *RT* leads the model to predict differences in *RT*s when classified according to the recognition judgment. We have described these predictions in detail elsewhere (see Berry et al., 2012, 2014), and only outline them here because they are tested in the experiments to follow. First, because the mean *f* of items judged old (i.e., items with $J_r > C$) tends to be greater than of items judged new (i.e., those with $J_r < C$), and because differences in mean *f* tend to translate to differences in *RT* (Equation 2), items judged old will tend to have shorter *RT*s than items judged new. This will be the case, even within old and new item types. The model therefore predicts that *RT*s to hits (old items judged old) will be faster than those of misses (old items judged new) (i.e.,

mean(*RT*| hit) < mean(*RT*| miss), and that the mean *RT* of false alarms (new items judged old) will tend to be faster than those of correct rejections (new items judged new) (i.e., mean(*RT*| false alarm) < mean(*RT*| correct rejection)).

Second, the model makes a prediction concerning the magnitude of the priming effect relative to the magnitude of the priming effect when calculated using only items judged new (i.e., misses and correct rejections). Although prior studies have taken priming for items judged new as evidence for the independence of the memorial sources driving recognition and priming (Stark & McClelland, 2000), this pattern is, in fact, predicted by the single system model simply because the mean *f* of misses tends to be greater than of correct rejections (when $\mu_{old} > 0$), which then translates to differences in *RT* (Equation 2) (see Berry et al., 2008). The mean difference in *f* between misses and correct rejections, however, will tend to be smaller than the mean difference in *f* between all new and old items. This is because the magnitude of the difference in J_r for misses and correct rejections is necessarily restricted by *C*. Differences in J_r tend to translate into differences in *f*, which in turn translate into differences in *RT*. The difference in *RT* to misses and correct rejections (i.e., the priming effect for items judged new) will therefore tend to be smaller than the difference in *RT* to new and old items (i.e., the priming effect).

In sum, the single-system model predicts that (1) experimental variables will not have opposite effects on recognition and priming because the magnitude of recognition and priming are both positively related to the strength of the memory signal in the model, so if a longer study duration improves recognition, then it will also lead to greater priming, though not necessarily by the same magnitude; (2) RTs of items judged old will be shorter than those of items judged new, even within old and new item types (i.e., mean(*RT*| hit) < mean(*RT*| miss), and also mean(*RT*| false alarm) < mean(*RT*| correct rejection)); (3) the priming effect will be greater than the priming effect for items judged new. Our primary concern in this

article is to test Prediction 1 by considering the effect of study duration on recognition and priming, though the nature of the paradigm used also enables Predictions 2 and 3 to be tested, potentially allowing for further validation of the model.

Experiments 1a and 1b

The aim of Experiment 1a was to reproduce the finding that study duration has opposite effects on recognition and priming, thereby providing data that would be challenging for the single-system model. At study, participants classified objects as natural or manufactured, which were presented for a relatively brief (250 ms) or long duration (2250 ms). In the test phase, on each trial they again classified old and new objects as natural or manufactured, before making a recognition judgment in relation to the item. The design of Experiment 1a was similar to Voss and Gonsalves' (2010) experiment but differed in the following ways: items in the long duration condition were presented for 2250 ms, rather than 2000 ms, there were fewer trials per study duration condition (48 vs. 85), there was a greater variation in the size of the stimuli (3-7° of visual angle vs. 4°), the images were drawn from a smaller pool (192 v. 320); participants' handedness was not controlled for in Experiment 1a, but all participants were right-handed in Voss and Gonsalves (2010); speed of responding was heavily emphasised in the test phase of Voss and Gonsalves (2010), but in Experiment 1a participants were instructed to make their classification response as quickly as possible but to not sacrifice accuracy; and finally, trials in the test phase were self-paced in Experiment 1a, whereas in Voss and Gonsalves' (2010) experiment each test trial advanced automatically. These differences were controlled for in Experiment 4, which used the same methods as Voss and Gonsalves (2010) (and, to preview, yielded a similar pattern of findings to Experiment 1a).

Experiment 1b used the same procedure as Experiment 1a except that, in the test phase, the classification task was replaced with a continuous identification with recognition

task (CID-R task, Feustel, Shiffrin, & Salasoo, 1983; Stark & McClelland, 2000) to measure recognition and priming. In this task, an item gradually clarifies from a mask on each test trial and participants press a button when they are able to identify the item; they then make a recognition judgment to the item. Repetition priming is evident if identification RTs to old items are shorter than those for new items.

The CID-R task was used in place of the classification task in Experiment 1b for several reasons: first, we sought to determine whether the priming results would generalise to a task in which the responses at test were not identical to the responses required at study, as is the case when the same classification task (natural vs. manufactured) is used in both the study task and also the task used to measure priming. When the response is identical at study and test, priming effects can be largely driven by retrieval of bindings between the stimulus (e.g., table) and the response (e.g., manufactured) that are formed during the study phase (Horner & Henson, 2009). This contrasts with the commonly held notion that priming effects can be driven by facilitation in the perceptual or conceptual processing of a stimulus. Any priming effects that occur in the CID-R task would arguably not be due to the retrieval of stimulus-response bindings because the type of response required to items at test (i.e., identification) is distinct from the type of response required in the study phase (i.e., classification) (Henson, Eckstein, Waszak, Frings, & Horner, 2014). Second, there is evidence to suggest that priming in the CID-R task is unaffected by intentional (explicit) retrieval strategies (Ward et al., 2013b), which is an important concern when measuring priming or implicit memory more generally (MacLeod, 2008). Finally, recognition and priming in the CID-R task have been shown to follow the predictions of the single-system model, and so if study duration has opposite effects on recognition and priming when measured with the CID-R task, then this would require re-evaluation of this earlier work.

Method

Participants.

There were 32 participants in Experiment 1a (classification task) (24 female, mean age = 20.8 years, *SD* = 3.2 years) and 24 participants in Experiment 1b (CID-R task) (23 female, mean age = 19.8 years, *SD* = 1.8 years). All individuals in this and subsequent experiments were psychology undergraduate students from University College London, the University of York, and/or Middlesex University who participated in partial fulfilment of a course requirement. Since the results from the classification task were more important for the purposes of replicating Voss and Gonsalves (2010), a greater number of participants were assigned to this task in this and subsequent experiments to ensure relatively high power. The power of Experiments 1a and 1b to detect an effect of study duration on priming of the same size as was found by Voss and Gonsalves (2010) (d_z = 0.855) was .997 and .980 (two-tailed), respectively.

Materials.

The stimuli were 192 color images of familiar objects (e.g., a tree, a chair) presented on a white background. The visual angle of each object ranged between 3 and 7 degrees in both the horizontal and vertical dimensions. Half were naturally occurring objects, and half were manufactured. For each participant, forty-eight images were randomly assigned to the 250 ms and 2250 ms study exposure duration conditions. These items were presented at study and test. The remaining 96 images were assigned to the new condition and were only presented at test. For each participant, images were randomly assigned to the 250 ms, 2250 ms, and new conditions such that each condition contained an equal number of natural and manufactured images. A different random assignment of images was used for each participant. The mask measured 12 x 12 degrees of visual angle and consisted of a 30 x 30 grid, the elements of which were randomly filled with fragments of pictures of objects that were not part of the main stimuli set.

Procedure.

Experiment 1a.

For the study phase, participants were told that they would see a series of objects in rapid succession and that they must press the "1" key or the "4" key to indicate whether they thought each object presented was 'natural' or 'manufactured (manmade)', respectively. They were told that the pictures would be presented quickly, for different durations, and that they should respond based on the first impression regarding the category. The labels "1 = natural", "4 = manufactured" were shown in 28 pt black Arial font below the central region where the stimuli appeared, and remained on screen as reminders for the duration of the study phase. Before the first trial commenced, the cue "Get Ready!" was presented in the centre of the screen in black 28 pt Arial font for 4 s. On each study trial, an image was presented for 250 ms or 2250 ms and was followed by the presentation of the mask for 2250 ms or 250 ms, respectively, such that the duration of each study trial was always 2500 ms. The same mask was used on each trial. The next trial was presented immediately following the presentation of the mask, which gave the appearance of a continuous sequence of flashing pictures, alternating with a mask. Voss and Gonsalves (2010) also presented trials in a continuous stream at study. Trials from the 250 ms and 2250 ms conditions were randomly intermixed. For a given trial, the response recording window lasted from the onset of the stimulus to the offset of the mask.

After the study phase, there was a retention interval of approximately 3 minutes, during which the participant read instructions for the test phase. These informed participants that, similar to the first stage, they would have to classify each object according to whether it is natural or manufactured by pressing 1 or 4, respectively, and to try to be as fast as they could when making their judgment, but not to sacrifice accuracy. At the start of each trial, a "+" fixation point was presented in the centre of the screen for 500 ms with the labels "1 =

natural, 4 = manufactured" presented beneath the stimulus presentation area, as in the study phase. An item was then presented for 500 ms, followed by the mask for 1500 ms. The categorization labels remained on screen until the participant made a response. Following a classification or identification response, as in Voss and Gonsalves (2010), the prompt for the recognition decision ("1 = *high confidence new*, 2 = *low confidence new*, 3 = *low confidence old*, 4 = *high confidence old*") was presented in the centre of the screen until a response was made. The initial instructions informed participants that "old" means that the item was presented in the first stage, and "new" means that it had not been presented before and that half of the objects were old. They were told that for this rating, it was more important to be accurate than fast. There were 192 trials in total, containing all items from the 250 ms, 2250 ms, and new conditions, randomly intermixed. Participants were tested individually, and the total duration of the experiment was approximately 30 minutes.

For each participant, individual RTs less than 200 ms were excluded from the analysis. The RT mean and *SD* were then calculated (collapsed across all conditions) and trials were removed from the analysis if the RT on the trial was greater than three times the standard deviation from the mean RT. This was done separately for the study and test phases. A correct categorization response to an item was defined as the modal categorisation judgment to the item at study and test across participants. The data were also analysed when trials were excluded on the basis of the correctness of classification responses, and the same pattern of results was found.[1]

In this and subsequent experiments, an alpha level of .05 was used for all statistical tests, and *t* tests are two-tailed, unless indicated. The Greenhouse-Geisser correction was applied on tests involving repeated-measures factors with more than two levels. The priming effect was calculated for each old item condition as the mean RT for new items minus the mean RT for old items. Cohen's d_z is given as the measure of effect size for the effect of

study duration on recognition and priming. This is calculated as $d_z = M_z / SD_z$ where the subscript *z* denotes the difference in scores between two measures in a within-subjects design. For repeated measures ANOVA, partial eta squared η_p^2 is also given as the effect size (as calculated by SPSS). For the analysis of the recognition data in this experiment and subsequent ones, ratings 1 and 2 were collapsed for "new" judgments and ratings 3 and 4 were collapsed for "old" judgments. The hit rate was then calculated as the proportion of old items judged old. The false alarm rate was calculated as the proportion of new items judged old. Responses were not analysed according to recognition confidence ratings in Experiments 1-3 because these results would not be directly relevant to our aims, and there are fewer stimuli than in Voss and Gonsalves' (2010) experiment. Finally, in order to test Predictions 2 and 3 of the single-system model, RTs to items at test were also analysed according to whether the recognition response was a hit, a miss (a "new" judgment to an old item), a false alarm or a correct rejection (a "new" judgment to a new item).

Experiment 1b.

The materials and procedures used in Experiment 1b were the same as those of Experiment 1a, except that priming for each item at test was measured using the CID procedure, rather than a classification task. The instructions informed the participant that a picture would be presented on each trial, and that it would become easier to identify with time; their task was to press the enter key as soon as they were sure that they could correctly identify the object. The sequence on each trial was as follows: the mask was presented for 500 ms. The picture was then presented for 17 ms (1 screen refresh at 60 Hz), and the mask then followed for 233 ms, forming a 250 ms presentation block. The stimulus was then immediately presented again for 17 ms, followed again by the mask for 233 ms. The picture was then presented for a slightly longer duration of 33 ms in the next presentation block, with the mask being presented for the remaining 217 ms of the block. Presentation continued in

this way, with the duration of the stimulus increasing by 17 ms on each alternate block and the mask being presented for the remainder of the 250 ms block. This gave the appearance of a stimulus that appeared to gradually clarify from a background mask. If the enter key had not been pressed after thirty 250 ms presentation blocks (i.e., after 7.5 s, when the stimulus duration equalled the block duration) then the cue "Please try to be faster on the next trial" appeared for 1s and the programme automatically advanced to the next test trial. If the participant pressed the enter key during the clarification procedure, a white outlined box was presented, into which the participant typed their identification response (e.g., 'tree'); they then pressed the enter key and were prompted for their recognition judgment using the same procedure as Experiment 1a.

Results: Experiment 1a

A small proportion of trials from the study phase could not be analysed because no key press response was made during the trial (M = 1.40%, SD = 2.72). Of the remaining study trials, a small proportion were removed because the RT was an outlier (M = 2.08%, SD = 1.03; see *Methods*). A small proportion of test phase trials were also excluded because the RT was an outlier (M = 1.53%, SD = 0.69; see *Methods*).

The main findings were that the priming effect did not differ between the 250 ms (M = 34 ms, SE = 11) and 2250 ms conditions (M = 34 ms, SE = 10), $t(31) = 0.004$, $p = .997$, d_z = 0.0006 (Table 1 and Figure 1a), whereas the recognition hit rate was significantly greater in the 2250 ms than the 250 ms condition, $t(31) = 6.29$, $p < .001$, $d_z = 1.11$ (Table 2 and Figure 1b). Seventeen out of 32 participants showed greater priming in the 2250 ms condition than the 250 ms condition, and 28 out of 32 participants had a higher hit rate in the 2250 ms condition than the 250 ms condition. Priming effects were reliable (i.e., greater than 0 ms) in the 250 ms condition, $t(31) = 3.02$, $p = .005$, and also the 2250 ms condition, $t(31) = 3.50$, p = .001. The hit rate was also significantly greater than the false alarm rate in both the 250 ms,

$t(31) = 34.08$, $p < .001$, and 2250 ms, $t(31) = 40.16$, $p < .001$, conditions, demonstrating that these items could be reliably discriminated from new items.

The RTs to items in the 250 ms and 2250 ms conditions did not significantly differ in the study phase, $t(31) = 1.15$, $p = .26$ (Table 3). The percentage of errors made in the classification task also did not significantly differ between the 250 ms and 2250 ms conditions at study, $t(31) = 0.15$, $p = .88$ (Table 4), or across the new, 250 ms, and 2250 ms conditions at test, $F < 1$ (Table 5). This suggests that the priming effects observed did not reflect a speed-accuracy trade-off.

Classification RTs at test were analysed according to the recognition judgment in order to test Predictions 2 and 3 of the single-system model. Participants in this and subsequent experiments were excluded from this analysis if they had zero responses in at least one of the hit, miss, false alarm or correct rejection categories. By this criterion, one participant was excluded from this analysis. As shown in Figure 2a, RTs to hits did not significantly differ from misses in either the 250 ms condition, $t(30) = -0.099$, $p = .92$, or the 2250 ms condition, $t(30) = -0.18$, $p = .86$. Similarly, RTs to false alarms did not significantly differ from correct rejections, $t(30) = 1.27$, $p = .22$; the effect ($M = 23$ ms, 95% CI [-14, 60]) only very weakly supported Prediction 2. The priming effect for items judged new (i.e., the mean RT for correct rejections minus the mean RT for misses) was not significant in the 250 ms condition ($M = 35$ ms, $SE = 18$), $t(30) = 1.91$, $p = .066$, or the 2250 ms condition, $t(30) = 1.65$, $p = .11$ ($M = 38$ ms, $SE = 23$). The priming effect did not differ from the priming effect for items judged new in either the 250 ms, $t(30) = -0.03$, $p = .98$, or 2250 ms conditions, $t(30) = -0.27$, $p = .79$, failing to support Prediction 3.

Results: Experiment 1b

A small proportion of trials from the study phase were excluded from the analysis because no key press response was made during the trial ($M = 0.87\%$, $SD = 1.36$). A further

M = 2.46% (*SD* = 1.48) of study trials were excluded because the RT was an outlier. For the analysis of the test phase responses, a small proportion of trials were not analysed because no identification keypress was made (*M* = 0.35%, *SD* = 0.68), and a further *M* = 1.83% (*SD* = 1.46) of trials were excluded because the RT was an outlier. The proportion of trials excluded because the item was incorrectly identified was similarly low (*M* < 1.5% of trials in each stimulus condition, Table 5), and there was no significant difference in the error rate across conditions, $F < 1$.

The magnitude of the priming effect was significantly greater in the 2250 ms (*M* = 352 ms, *SE* = 40) condition than the 250 ms condition (*M* = 257 ms, *SE* = 27), $t(23) = 2.26$, $p = .034$, $d_z = 0.46$ (greater in 15 out of 24 participants) (Table 1 and Figure 3a). The hit rate was also greater in the 2250 ms condition than the 250 ms condition, $t(23) = 5.44$, $p < .001$, $d_z = 1.11$ (in 21 out of 24 participants) (Table 2 and Figure 3b). The priming effect was significant in both the 250 ms, $t(23) = 9.45$, $p < .001$, and 2250 ms, $t(23) = 9.35$, $p < .001$, conditions. In the recognition task, the hit rate was significantly greater than the false alarm rate in both the 250 ms, $t(23) = 15.85$, $p < .001$, and 2250 ms, $t(23) = 19.48$, $p < .001$, conditions. These priming and recognition results are therefore in line with Prediction 1 of the single-system model. In the study phase, classification RTs did not significantly differ between the 250 ms and 2250 ms conditions, $t(23) = 0.37$, $p = .72$ (Table 3), nor did the percentage of incorrect classification responses differ between conditions, $t(23) = 1.96$, $p = .063$ (Table 4).

When analysed according to the recognition memory judgement, RTs to hits were found to be significantly shorter than RTs to misses in the 250 ms condition, $t(23) = 2.97$, $p = .007$, and the 2250 ms condition, $t(23) = 3.97$, $p = .001$ (Figure 2b). RTs to false alarms were also significantly faster than those of correct rejections, $t(23) = 2.89$, $p = .008$. This confirmed Prediction 2 of the model. Furthermore, the priming effect for items judged new

was significant (i.e., > 0 ms) in the 250 ms condition, $t(23) = 2.65$, $p = .014$ ($M = 150$ ms, $SE = 57$), but this was not the case for misses in the 2250 ms condition, $t(23) = 0.76$, $p = .46$ ($M = 76$ ms, $SE = 101$). Finally, the priming effect was significantly greater than the priming effect for items judged new in the 2250 ms condition, $t(23) = 3.45$, $p = .002$, but not the 250 ms condition, $t(23) = 1.85$, $p = .078$, offering some support for Prediction 3.

Discussion

In Experiments 1a and 1b, no evidence was found to support the claim that recognition and priming are affected in opposite ways by study duration. In both experiments, recognition memory was better for items presented for 2250 ms at study than 250 ms. Priming did not differ between these conditions when measured using a classification task (Experiment 1a), but when measured using the CID-R task, the magnitude of priming was significantly greater in the 2250 ms condition than the 250 ms condition, which is contrary to Voss and Gonsalves' (2010) finding. Thus, Prediction 1 of the model was confirmed with the CID-R task, but not the classification task. Support for Predictions 2 and 3 was also found with the CID-R task, but not with the classification task; we discuss this further in sections *Discussion of Experiments 1-3, Modelling of Experiments 1-4* and the *General Discussion*.

Experiments 2a and 2b

Given the difference in results of Experiments 1a and 1b to those of Voss and Gonsalves (2010), the aim of Experiments 2a and 2b was to replicate the results of Experiments 1a and 1b while improving upon one aspect of the study phase procedure. The rapid, continuous nature of the trial presentation in the study phase of Experiments 1a and 1b meant that some objects were not classified within the 2.5 s trial duration.[2] That is, in some instances, a new trial was presented before participants had provided a classification response in relation to the object. It is possible that this may have introduced an undesired unsystematic source of variability, or noise, into the way in which items on some trials were

encoded. In an extreme case, if a participant happened to be particularly slow to respond on a given study trial, a classification response to an item may have been made during the stimulus presentation on the following trial. In order to rule out the possibility that this noise in encoding impacted upon the ability to detect differences in priming between the 250 ms and 2250 ms conditions, an identical procedure to Experiments 1a and 1b was used in Experiments 2a and 2b except that on each study trial, a classification response had to be made to an item before the next trial was presented, thereby ensuring that every item was overtly classified in the study phase.

Method

Participants.

There were 32 participants in Experiment 2a (26 female, mean age = 20.8 years, *SD* = 3.2 years) and 18 participants in Experiment 2b (6 female, mean age = 20.6 years, *SD* = 1.7 years).

Materials and procedure.

Experiments 2a and 2b were identical to Experiments 1a and 1b except that the study phase procedure was modified so that a classification response was collected for every item in the study phase. At the start of each study trial a "+" was shown for 500 ms. The stimulus was then presented for 250 ms or 2250 ms, and the mask was then presented for 2250 ms or 250 ms, followed by a blank screen. If no response had been made by the time the mask presentation had terminated then the programme waited for the participant to make a classification response before advancing. There was an inter-trial interval of 750 ms during which a blank screen was presented.

Results for Experiment 2a

The proportion of study trials that were removed because the RT was an outlier was low (M = 1.92%, SD = 1.09). Similarly, the proportion of test trials that were excluded because the RT was an outlier was low (M = 1.68%, SD = 0.89).

As found in Experiment 1a, the magnitude of priming did not significantly differ between the 250 ms (M = 44 ms, SE = 10) and 2250 ms (M = 64 ms, SE = 13) study exposure duration conditions, $t(31) = 1.70$, $p = .10$, $d_z = 0.30$ (Figure 1c and Table 1), although the mean priming effect was numerically greater in the 2250 ms condition (shown by 19 out of 32 participants). Also, as in Experiment 1a, the recognition hit rate was significantly greater in the 2250 ms condition than the 250 ms condition, $t(31) = 9.46$, $p < .001$, $d_z = 1.67$ (Figure 1d and Table 2) (shown by 29 out of 32 participants). Priming was reliable (i.e., greater than 0 ms) in both the 250 ms, $t(31) = 4.50$, $p < .001$, and 2250 ms, $t(31) = 4.84$, $p < .001$, conditions. The hit rate was also significantly greater than the false alarm rate in both the 250 ms, $t(31) = 26.24$, $p < .001$, and 2250 ms, $t(31) = 30.14$, $p < .001$, conditions, indicating that participants could discriminate old from new items.

In the study phase, there was no significant difference between RTs to items in the 250 ms and 2250 ms conditions, $t(31) = 0.97$, $p = .34$ (Table 3), nor was there a significant difference between the percentage of classification errors between conditions, $t(31) = 0.49$, $p = .63$ (Table 4). Unlike in Experiment 1a, there was a significant difference between the percentage of classification errors to items in the new, 250 ms, and 2250 ms conditions at test, $F(1.99, 61.58) = 6.73$, $p = .002$ (Table 5). This was primarily driven by the difference in error rates between the new and 250 ms conditions ($p < .001$), and there was no reliable difference in the error rates between the new and 2250 ms conditions ($p = .12$) or the 250 ms and 2250 ms conditions ($p = .057$). Given that the difference in error rates to the 250 ms and new items was unexpected and was not found in any other experiment, it is not given further consideration.

Classification RTs at test were analysed according to the recognition outcome in order to test Predictions 2 and 3 of the single-system model. One participant was excluded from this analysis for having zero responses in at least one of the hit, miss, false alarm, or correct rejection categories. As in Experiment 1a, RTs to hits did not significantly differ from misses in either the 250 ms, $t(30) = 1.28$, $p = .11$, or 2250 ms, $t(30) = 0.86$, $p = .40$, conditions (see Figure 2c); the effects (M = 28 ms, 95% CI [-17, 74], and M = 27 ms, 95% CI [-36, 90], respectively) only weakly supported Prediction 2 . RTs to false alarms also did not differ from those of correct rejections, $t(30) = -0.10$, $p = .93$, which does not support Prediction 2. Priming for items judged new was not reliable in either the 250 ms condition, $t(30) = 0.75$, $p = .46$ (M = 13 ms, SE = 17), or the 2250 ms condition, $t(30) = 1.06$, $p = .30$ (M = 31 ms, SE = 29). The priming effect did not differ from the priming effect for items judged new in either the 250 ms, $t(30) = 1.68$, $p = .10$, or 2250 ms conditions, $t(30) = 0.92$, $p = .37$; the effects (M = 26 ms, 95% CI [-6, 57] and M = 26 ms, 95% CI [-32, 84], respectively) only weakly supported Prediction 3.

Results for Experiment 2b

A small proportion of trials were excluded from the analysis of the study phase because the RT was an outlier (M = 1.50%, SD = 0.89). The proportion of trials that were excluded from the analysis of the test phase because no key press was made before the termination of the clarification procedure was low (M = 2.00%, SD = 3.86). A further M = 3.03% (SD = 4.45) of the remaining trials at test were excluded because the RT was an outlier. The number of trials excluded because the item was incorrectly identified at test was also low (M < 5% in each condition, Table 5) and did not significantly differ between new, 250 ms, and 2250 ms conditions, $F < 1$.

As found in Experiment 1b, priming was significantly greater in the 2250 ms (M = 399 ms, SE = 56) condition than the 250 ms condition (M = 264 ms, SE = 43), $t(17) = 2.60$, p

= .019, d_z = 0.61 (and greater in 12 out of 18 participants) (Figure 3c and Table 1). Likewise, the hit rate was significantly greater in the 2250 ms condition than the 250 ms condition, $t(17) = 2.15$, $p = .047$, $d_z = 0.51$ (being greater in 12 out of 18 participants) (Figure 3d and Table 2). The priming effect was reliable (greater than 0 ms) in both the 250 ms, $t(17) = 6.16$, $p < .001$, and 2250 ms, $t(17) = 7.08$, $p < .001$, conditions. The hit rate was greater than the false alarm rate in both the 250 ms, $t(17) = 7.35$, $p < .001$, and 2250 ms conditions, $t(17) = 8.12$, $p < .001$. In the study phase, there was no significant difference in classification RTs between the 250 ms and 2250 ms conditions (Table 3), $t(17) = 0.81$, $p = .43$, nor was there a significant difference in the percentage of classification errors (Table 4), $t(17) = 1.16$, $p = .26$.

When RTs were analysed according to the recognition outcome, RTs to hits significantly differed from misses in the 2250 ms condition, $t(17) = 2.19$, $p = .043$, supporting Prediction 2 of the model, but this was not the case in the 250 ms condition, $t(17) = 1.34$, $p = .20$ (Figure 2d); the effect in this condition (M = 108 ms, 95% CI [-62, 279]) only very weakly supported Prediction 2. RTs to correct rejections also did not significantly differ from those of false alarms, $t(17) = 0.39$, $p = .70$; the effect (M = 59 ms, 95% CI [-261, 378]) only weakly supported Prediction 2. Thus, there was only partial support for Prediction 2. Priming for items judged new was significant (i.e., > 0 ms) in both the 250 ms, $t(17) = 3.65$, $p = .001$ (M = 233 ms, SE = 64), and 2250 ms, $t(17) = 3.03$, $p = .01$ (M = 239 ms, SE = 79), conditions. Finally, despite numerical trends, the priming effect was not significantly greater than the priming effect for items judged new in either the 2250 ms condition, $t(17) = 2.01$, $p = .061$, or the 250 ms condition, $t(17) = 0.51$, $p = .62$; the effects (M = 160 ms, 95% CI [-8, 328] and M = 30 ms, 95% CI [-96, 156], respectively) only weakly supported Prediction 3..

Discussion

The recognition and priming results of Experiments 2a and 2b replicated those of Experiments 1a and 1b, but with a refined study procedure that ensured that every item was

classified at study. Recognition was greater for items in the 2250 ms condition than the 250 ms condition; priming did not reliably differ between the 250 ms and 2250 ms conditions in the classification task (Experiment 2a), but was significantly greater in the 2250 ms condition than the 250 ms condition when measured using a CID-R task (Experiment 2b). Thus, as in Experiments 1a and 1b, Prediction 1 of the single-system model was confirmed with the CID-R task, but not the classification task. Predictions 2 and 3 were not confirmed in the classification task. There were numerical trends in support of Predictions 2 and 3 in the CID-R task, but only one result concerning Prediction 2 was significant.

Experiments 3a and 3b

The aim of Experiments 3a and 3b was to address the possibility that opposite effects of study duration on recognition and priming might become evident when a wider range of study durations is used than in Experiments 1a, 1b, 2a, and 2b. This is important to address for two reasons: first, the duration of the long study duration condition in Experiments 1 and 2 (2250 ms) differed from the duration of the long condition used by Voss and Gonsalves (2010) (2000 ms), and it is possible that Voss and Gonsalves' results would have been replicated had a duration closer to 2000 ms been used in the long condition. Second, Zago et al. (2005) found evidence to suggest that the function relating exposure duration to priming can be inverse U-shaped. They found that priming increased as study duration increased from 40 ms to 250 ms and then fell as study duration increased further to 1900 ms. If it existed, this kind of inverse U-shaped relationship would, of course, be missed in our previous experiments because only two study duration conditions were used. It could be argued that consideration of a greater number of study durations would allow for the true (non-montonically increasing) nature of the function relating study duration to priming to be revealed, and that this function may not be the same for recognition. Accordingly, in order to determine whether evidence for a dissociation would be found when a wider range of study

durations was used, Experiments 3a and 3b were very similar to Experiments 1a and 1b except that six different study durations were used (i.e., 40 ms, 150 ms, 250 ms, 350 ms, 500 ms, and 1900 ms, as were used in Zago et al., 2005).

Method

Participants.

There were 30 participants in Experiment 3a (28 female, mean age = 20.5 years, *SD* = 3.0 years) and 30 participants in Experiment 3b (27 female, mean age = 20.3 years, *SD* = 3.5 years). One participant from Experiment 3a was excluded from the analysis for making a high proportion of classification errors during the test phase (67% errors). Another participant was excluded from the analysis of Experiment 3b for failing to follow instructions in the study phase (no key press responses were recorded).

Materials and procedure.

The stimuli were 425 pictures of objects, taken from the same set of stimuli used by Zago et al. (2005); each subtended approximately 3 degrees of visual angle in the horizontal and vertical dimensions. There were six study duration conditions (40 ms, 150 ms, 250 ms, 350 ms, 500 ms, or 1900 ms) and a new item condition. For each participant, sixty items were randomly assigned to each of these conditions. A further five items were used on practice trials for the study phase. There were 10 masks, which were randomly assigned to the study trials with the constraint that each of the 10 masks appeared an equal number of times across study trials. The stimulus-plus-mask duration was always 2000 ms for each trial (as in Zago et al., 2005). Trials were presented in a continuous stream as in Experiment 1a and Zago et al. (2005). The test phase procedure of Experiment 3a was the same as Experiments 1a and 2a and all studied and new items were presented in this phase (totalling 420 trials). Trials were arranged into 6 blocks of equal length, and an equal number of new trials were presented in each block.

The materials and procedure in Experiment 3b were identical to those in Experiment 3a except that the CID-R task was used to measure priming at test and thirty items, rather than sixty were randomly assigned to each of the 40 ms, 150 ms, 250 ms, 350 ms, 500 ms, and 1900 ms conditions. Fewer items were used because of the tendency for CID trials to be longer, and this prevented this phase from being overly long, relative to the classification task. There were therefore 210 trials in total at test in Experiment 3b.

Results for Experiment 3a

A small proportion of trials were excluded from analysis of the study phase because no key press response was made during the trial (M = 6.62%, SD = 6.27). Of the remaining study trials, a small proportion were excluded because the RT was an outlier (M = 2.96%, SD = 2.43). The proportion of trials that were excluded from analysis of the test phase because the RT was an outlier was similarly low (M = 1.84%, SD = 1.49).

The magnitude of priming tended to increase across the 40 ms, 150 ms, 250 ms, 350 ms, 500 ms, and 1900 ms conditions, as indicated by a significant repeated-measures analysis of variance (ANOVA), $F(3.96, 110.91) = 3.52$, $p = .01$, $\eta_p^2 = .11$, and significant linear trend, $F(1,28) = 13.67$, $p = .001$ (Table 1 and Figure 1e). Focussing on the 250 ms and 1900 ms conditions in particular, because these two conditions are the most similar to those of Experiments 1a and 1b, the priming effect was significantly greater in the 1900 ms condition (M = 72 ms, SE = 19) than the 250 ms condition (M = 36 ms, SE = 18), $t(28) = 2.21$, $p = .036$, $d_z = 0.41$ (shown by 18 out of 29 participants). Priming was also reliable (i.e., greater than 0 ms) in all conditions, ts > 2.48, ps < .02, except in the 40ms condition, $t(28) = 1.17$, $p = .27$, and was only marginally significant in the 250 ms condition, $t(28) = 2.02$, $p = .053$.

With regards to the recognition data, the hit rate tended to increase across the 40 ms, 150 ms, 250 ms, 350 ms, 500 ms, and 1900 ms conditions, $F(2.13, 59.63) = 86.44$, $p < .001$, $\eta_p^2 = .76$ (Table 2 and Figure 1f). The linear trend was also significant, $F(1, 28) = 141.12$, $p <$

.001, η_p^2 = .83. The hit rate was significantly greater than the false alarm rate in all study duration conditions (*t*s > 3.22, *ps* < .003), indicating that recognition memory was reliable in all conditions. As with the priming data, the hit rate for the 1900 ms condition was greater than that of the 250 ms condition, $t(28) = 6.82$, $p < .001$, $d_z = 1.27$ (in 27 out of 29 participants). Thus, in this experiment, Prediction 1 of the model was confirmed in the classification task.

There was a significant difference in the percentage of classification errors made in the 40 ms, 150 ms, 250 ms, 350 ms, 500 ms, or 1900 ms conditions at study, $F(2.26, 63.35) = 39.66$, $p < .001$ (Table 4). This appeared to be driven by the high percentage of errors in the 40 ms condition and when this condition was removed from the analysis the difference in the error rate in the remaining conditions was not significant, $F < 1$. Similarly, there was a significant difference in the RTs to items across conditions in the study phase, $F(2.02, 56.47) = 10.22$, $p < .001$, but this reflected the longer RTs in the 40 ms condition and again this difference was not significant when the 40 ms condition was excluded from the analysis, $F < 1$ (Table 3). In the test phase, there was no significant difference in the percentage of classification errors made in the new, 40 ms, 150 ms, 250 ms, 350 ms, 500 ms, and 1900 ms conditions, $F(3.86, 108.05) = 0.09$, $p = .99$ (Table 5).

When RTs were analysed according to the recognition outcome, no significant differences were found between the items judged old and new within the new, 40 ms, 150 ms, 250 ms, 350 ms, 500 ms or 1900 ms conditions (*t*s < 1.91, *p*s > 0.067) (Figure 2e). The effects in the 40 ms (*M* = -11 ms, 95% CI [-53, 31]) and new (*M* = 0 ms, 95% CI [-55, 55]) conditions did not support Prediction 2, whereas the effects in the 150 ms (*M* = 27 ms, 95% CI [-24, 77]), 250 ms (*M* = 45 ms, 95% CI [-3, 94]), 350 ms (*M* = 50 ms, 95% CI [-5, 106]), 500 ms (*M* = 29 ms, 95% CI [-24, 81]), and 1900 ms (*M* = 21 ms, 95% CI [-35, 77]) conditions weakly supported Prediction 2. There was no significant difference between RTs

to correct rejections and misses in each of the old item conditions (all *t*s < 1.88, *p*s > .07), and the priming effect did not significantly differ from the priming effect for items judged new in any of the conditions (*t*s < 1.71, *p*s > .10); the effect in the 40 ms condition (*M* = -7 ms, 95% CI [-25, 12]) did not support Prediction 3, whereas the effects in the 150 ms (*M* = 15 ms, 95% CI [-17, 46]), 250 ms (*M* = 27 ms, 95% CI [-8, 62]), 350 ms (*M* = 35 ms, 95% CI [-7, 78]), 500 ms (*M* = 24 ms, 95% CI [-13, 60]), and 1900 ms (*M* = 15 ms, 95% CI [-29, 59]) conditions only very weakly supported Prediction 3.

Results for Experiment 3b

A proportion of study trials could not be analysed because no key press response was made during the trial (*M* = 10.77%, *SD* = 7.54). Of the remaining study trials, a small proportion were not analysed because the RT was an outlier (*M* = 2.42%, *SD* = 1.19). For the test phase, the proportion of trials that were excluded from the analysis because no key press was made before the termination of the clarification phase was low (*M* = 3.40%, *SD* = 4.74; see *Method* section), as was the proportion of the remaining test trials that were excluded because the RT was an outlier (*M* = 0.92%, *SD* = 0.66). The number of test trials excluded because the item was incorrectly identified was very low (*M* < 2.5% in each condition; Table 5) and did not significantly differ between conditions, $F(4.58, 128.34) = 1.20$, $p = .31$.

The priming effect tended to increase across the 40 ms, 150 ms, 250 ms, 350 ms, 500 ms, and 1900 ms conditions (Figure 3e, Table 1), and this was confirmed by a significant repeated-measures ANOVA, $F(4.08, 114.12) = 13.05$, $p < .001$, $\eta_p^2 = .32$, and significant linear trend $F(1, 28) = 47.27$, $p < .001$, $\eta_p^2 = .32$. The quadratic and cubic trends were also significant (*F*s(1, 28) > 9.61, *p*s < .005, though see below). Priming was reliable (greater than zero) in the 150 ms, 250 ms, 350 ms, 500 ms, and 1900 ms conditions, *t*s > 4.95, *p*s < .001, but not in the 40 ms condition, $t(28) = 0.64$, $p = .53$. When the 40 ms exposure duration condition was excluded in the repeated-measures ANOVA, the overall ANOVA was no

longer significant, $F(3.53, 98.7) = 2.02$, $p = .11$, the linear trend was only marginally significant, $F(1, 28) = 3.70$, $p = .065$, and the quadratic and cubic trends were no longer significant, $F(1, 28) = 0.44$, $p = .51$, and $F(1,28) = 1.76$, $p = .20$, respectively. Focussing again on the 250 ms and 1900 ms conditions, although priming was numerically greater in the 1900 ms condition (M = 430 ms, SE = 63, vs. M = 385 ms, SE = 56), the difference was not significant, $t(28) = 0.87$, $p = .39$, $d_z = 0.16$ (only 14 out of 29 participants showed greater priming in the 1900 ms than the 250 ms condition).

In the recognition memory data, the hit rate also tended to increase across the 40 ms, 150 ms, 250 ms, 350 ms, 500 ms, and 1900 ms conditions, $F(2.59, 72.46) = 25.54$, $p < .001$, $\eta_p^2 = .48$ (linear trend: $F(1, 28) = 37.58$, $p < .001$, $\eta_p^2 = .57$) (Figure 3f, Table 2). The quadratic and cubic trend components were also significant (Fs(1,28) > 22.09, ps < .001; though see below). However, in the 40 ms condition, as with the priming data, the hit rate was not significantly greater than the false alarm rate, $t(28) = 0.60$, $p = .55$, which indicated that recognition memory was no better than chance for these items. The hit rate in the 150 ms, 250 ms, 350 ms, 500 ms, and 1900 ms conditions was, however, significantly greater than the false alarm rate (ts > 5.67, $ps < .001$). When the 40 ms condition was excluded from the analysis of differences in the hit rate between conditions, the difference between the 150 ms, 250 ms, 350 ms, 500 ms, and 1900 ms conditions remained significant, $F(3.12, 87.40) = 2.84$, $p = .04$, as did the linear trend, $F(1, 28) = 8.00$, $p = .009$, but not the quadratic, $F(1, 28) = 0.095$, $p = .76$, or cubic, $F(1, 28) = 1.21$, $p = .28$, trends. Again, comparing the 250 ms and 1900 ms conditions, although the hit rate in the 1900 ms condition was numerically greater than that of the 250 ms condition, the difference was not significant, $t(28) = 1.55$, $p = .13$, $d_z = 0.29$ (being greater for 17 out of 29 participants).

There was a significant difference in RTs to items at study across the 40 ms, 150 ms, 250 ms, 350 ms, 500 ms, and 1900 ms conditions, $F(2.76, 77.34) = 3.20$, $p = .031$ (Table 3).

As in Experiment 3a, this appeared to be largely driven by the generally longer RTs to items in the 40 ms condition, and the ANOVA was not significant when this condition was excluded, $F(2.27, 63.60) = 1.47$, $p = .24$. Similarly, there was also a significant difference in the percentage of classification errors made across conditions at study, $F(2.71, 75.76) = 27.76$, $p < .001$ (Table 4). Again, this was largely driven by the high percentage of errors in the 40 ms condition, and the ANOVA was not significant when this condition was removed from the analysis, $F(3.31, 92.78) = 2.51$, $p = .058$.[3]

For the analysis of RTs according to the recognition judgment, three participants were not included in this analysis because at least one of the hit, miss, false alarm or correct rejection response categories had zero responses. RTs were faster for items judged old in the new item condition, $t(25) = 2.48$, $p = .02$, and also the 40 ms, $t(25) = 3.39$, $p = .002$, 250 ms, $t(25) = 2.79$, $p = .01$, 350 ms, $t(25) = 2.15$, $p = .042$, and 1900 ms conditions, $t(25) = 2.26$, $p = .033$, offering some support for Prediction 2 of the model (as shown in Figure 2f). There was, however, no reliable difference in RTs to items judged old or new in the 150 ms, $t(25) = 0.93$, $p = .36$, or 500 ms, $t(25) = 0.46$, $p = .65$, conditions; the effect in these conditions (M = 111 ms, 95% CI [-135, 356], and M = 42 ms, 95% CI [-145, 229], respectively) only weakly supported Prediction 2 of the model.The priming effect for items judged new was significant in the 150 ms, $t(25) = 2.39$, $p = .025$, 250 ms, $t(25) = 2.75$, $p = .011$, 500 ms, $t(25) = 3.76$, $p = .001$, and 1900 ms, $t(25) = 2.54$, $p = .018$, conditions, but not in the 40 ms condition, $t(25) = 0.50$, $p = .62$, or 350 ms condition, $t(25) = 1.95$, $p = .063$. The only condition in which the priming effect was greater than the priming effect for items judged new (relevant to Prediction 3 of the model) was the 250 ms condition, $t(25) = 2.55$, $p = .017$ (all other ts < 1.75, ps > .093); the effects in the 40 ms (M = 75 ms, 95% CI [-18, 168]), 150 ms (M = 53 ms, 95% CI [-67, 172]), 350 ms (M = 109 ms, 95% CI [-59, 276]), 500 ms (M = 10 ms, 95%

CI [-113, 133]), and 1900 ms (M = 146 ms, 95% CI [-26, 317]) conditions only weakly supported Prediction 3.

Discussion

As in the previous experiments, no evidence was found in Experiments 3a and 3b to support the claim that varying study duration has opposite effects on recognition and priming. To the contrary, both recognition and priming significantly increased across the 40 ms to 1900 ms study duration conditions in the classification and CID-R tasks, supporting Prediction 1 of the model. In the classification task, priming for items in the longest study duration condition (1900 ms) was also significantly greater than priming for items presented for 250 ms (Experiment 3a). The lack of a significant difference between priming in the 1900 ms and 250 ms conditions using the CID-R task (Experiment 3b), contrasts with the findings of Experiments 1b and 2b where a significant difference was found. This may be due to lower power arising from the smaller number of stimuli used in Experiment 3b. The function relating priming and recognition to study duration was found to be a monotonically increasing one, rather than an inverse U shaped one reported by Zago et al. (2005); reasons for the difference in findings are considered in greater detail in the *General Discussion*. Finally, once again, Predictions 2 and 3 of the single-system model were not supported in the classification task, but there was some support for these predictions in the CID-R task.

Discussion of Experiments 1-3

It is important to consider potential reasons why the priming results from Experiments 1-3 so far did not replicate those of Voss and Gonsalves (2010). A power analysis was conducted to determine whether Experiments 1-3 possessed sufficient power to detect the priming difference reported in Voss and Gonsalves (2010). Using the results reported in Voss and Gonsalves, the effect of study duration on priming in their study was estimated to be relatively large (Cohen's $d_z = t/\sqrt{N} = 3.2/\sqrt{14} = 0.855$). The power of each individual

experiment was very high (all greater than .93, two-tailed test). For comparison, the effect of duration on the recognition hit rate was estimated to be only slightly larger (Cohen's $d_z = t/\sqrt{N} = 4.8/\sqrt{14} = 1.28$). Thus, low power is unlikely to be the cause of the failure to replicate the priming results of Voss and Gonsalves.

Although the design we used was similar, the methodological differences between our experiments and Voss and Gonsalves' may have contributed to the discrepancy in results. First, there were fewer items in each condition of our experiments. In the experiment by Voss and Gonsalves, there were 85 items in the brief and long conditions at encoding (and 150 new items at test), whereas there were 48 items in the 250 ms and 2250 ms conditions in Experiments 1a, 1b, 2a, and 2b, 60 items in the 250 ms and 1900 ms conditions of Experiment 3a, and only 30 in these conditions in Experiment 3b. The stimuli were also not identical to those used by Voss and Gonsalves: the stimuli in Experiments 1 and 2 were obtained from image searches on the internet, and those in Experiment 3 were taken from the study by Zago et al. (2005). The stimuli were also of different dimensions.

Another possible reason for the discrepancy in findings is that participants in Experiments 1-3 were generally slower to respond in the classification task compared to participants in Voss and Gonsalves' study. Pooing the data from Experiments 1a, 2a, and 3a, the mean RT to new items was 976 ms (*SD* = 313) whereas the mean RT to new items was 627 ms (*SD* =161) in Voss and Gonsalves' study.[4] Average RTs to new items were also faster in Zago et al. (2005) (696 ms and 770 ms in their block- and event-related designs, respectively). It could be argued that the generally longer RTs shown by participants in our study meant that there was more opportunity for their responses in the priming task to be contaminated by explicit memory. Explicit memory increased with study duration (indicated by recognition performance), and if contamination was occurring, priming might also be expected to increase with study duration. There are, however, a number of reasons to expect

that explicit contamination did not occur: first, the test phase instructions for the classification task encouraged rapid responding, as in Voss and Gonsalves' study, meaning that there would have been limited time for contamination to occur. Moreover, a speeded classification decision is arguably less cognitively demanding than a strategy based on recollection (e.g., Macleod, 2008), meaning that it is unlikely that participants would have opted to engage in such a strategy. It is true, however, that we instructed participants not to sacrifice accuracy when responding quickly at test, whereas Voss and Gonsalves (2010) did not, and this difference may be important. Second, even under circumstances designed to encourage contamination of priming responses by explicit memory in similar tasks (with interleaved identification and recognition trials), contamination of priming with the use of an explicit strategy has not been found (Ward et al., 2013b). Third, the idea that explicit contamination was more likely to have occurred in our experiments would seem more plausible if levels of recognition were greater than reported by Voss and Gonsalves (2010). This would indicate a greater availability of information in explicit memory for retrieval. Levels of recognition were, however, comparable to those obtained by Voss and Gonsalves (2010), and, if anything, in many experiments were slightly lower (as indicated by the hit rate minus the false alarm rate: .62 and .73 in the brief and long conditions of Voss and Gonsalves, respectively; see Table 2 for Experiments 1-3).

Fourth, if slower baseline RTs are responsible for the failure to detect an inverse relationship between study duration and priming, then we might expect to see evidence of an inverse relationship in responders with baseline RTs comparable to those of participants in Voss and Gonsalves (2010). This was examined by collapsing the data across experiments for each task and then partitioning the participants into quartiles on the basis of their mean RT to new items and then calculating the difference in priming to long and brief items in each quartile. The results are shown in Table 6. For the classification task data, the mean RT for

new items in the fastest quartile was comparable to (and was if anything faster than) the mean RT for new items reported in Voss and Gonsalves (*M* = 590 ms vs. *M* = 627 ms). Yet, there was virtually no difference in priming between long and brief conditions in this quartile (*M* priming long – *M* priming brief = 0.69 ms) and, importantly, there was no reversal in the priming difference here; priming was, if anything, greater in the long condition (12 out of 23 participants showed a positive difference between long and brief conditions). There was also no evidence for a reversal in the CID-R task in the fastest responders.

Clearly, however, in both tasks, the long-minus-brief-priming difference tended to decrease with the mean RT for new items, raising the possibility that if the mean RT of new items was even faster than 590 ms in the classification task, a negative long minus brief priming difference might emerge (e.g., in a direct replication that encouraged very fast responding). Although possible, we do not believe that the association between the duration effect and baseline RT necessarily implies it. For example, this pattern of results is consistent with the findings of Ostergaard (1998), who found that independent variables such as the number of repetitions at encoding, word-frequency, and retention interval had no effect on the magnitude of subsequent priming when baseline RTs were fast, but effects of these variables emerged when baseline RTs were longer. Rather than explain this in terms of the differential engagement of explicit and implicit memory systems as baseline RT varies, Ostergaard (1998) proposed that when baseline RTs were short, many factors unrelated to memory (e.g., the amount of perceptual information available from a stimulus at test) dominated responding in the priming task (a word naming task). This effectively constrained the effects of the independent variables on priming. When the influence of these factors was reduced (e.g., by lengthening the stimulus fade-in duration in a gradual clarification task, thereby increasing perceptual difficulty), there was more opportunity for the influence of prior exposure to facilitate performance in the priming task, and therefore effects of

independent variables on priming could be detected (see also Ostergaard, 1999, for a similar set of findings in amnesia). This account suggests that there would still be no difference in priming, even if baseline RTs were below 590 ms on average. In sum, this analysis suggests that the failure to find that study duration has opposite effects on priming and recognition in Experiments 1-3 (particularly Experiments 1a, 2a, and 3a) is unlikely to be due to differences in baseline RT between studies.

Experiment 4

Given that we cannot exclude the possibility that methodological differences to Voss and Gonsalves (2010) in Experiments 1-3 may have been important in driving the differences in results between studies, Experiment 4 was designed to replicate the behavioural methods of Voss and Gonsalves (2010) as closely as possible. With the exception of additional methodological controls and counterbalancing procedures (see below), the only potential substantive differences to the methods of Voss and Gonsalves (2010) that we can discern in this experiment are that the stimuli and participants are not the same.[5] The methods and analysis plan were registered prior to data collection (with the Open Science Framework, https://osf.io/xfqt6) (any unplanned analyses are indicated as such in parentheses).

Method

Participants.

40 individuals participated (37 female, 18-25 years of age). This particular sample size was selected because it is at least 2.5 times that of Voss and Gonsalves (2010) (N = 14), and so would give the experiment at least 80% power to detect an effect size that would give the original Voss and Gonsalves (2010) study 33% power (Simonsohn, 2015). This sample size also allowed for counterbalancing of stimuli and responses (see below). All participants were native English speakers, right-handed, and between 18-25 years of age. (In Voss and Gonsalves, the participants were right-handed and between 18-24 years of age, but whether

participants were native English speakers was not specified.) The participants were recruited from a University of Plymouth participant pool, and took part in exchange for course credit.

Materials.

The stimuli were 320 colour images. Each image consisted of a photograph of a nameable, familiar object against a white background. The images were selected from the stimulus pool provided by Brodeur, Guérard, and Bouras (2014) and the stimuli used by Zago et al. (2005). The objects in the stimulus pool were different from one another. Each image was sized to fit into a square subtending 4 degrees of the visual angle (from a viewing distance of approximately 50 cm). Half of the images were natural (e.g., an apple, wheat, a moth) and half were manufactured (e.g., a bed, a chair, a guitar). As in Voss and Gonsalves (2010), the assignment of stimuli to the brief and long conditions was counterbalanced across participants and approximately half of the items in each condition were manufactured. The mask was composed of random elements of objects not in the main stimulus pool and had the same size dimensions as the stimuli. A four-button response box with millisecond response time accuracy was used to collect button press responses.

Procedure.

On each trial in the study phase, an item was presented and was then immediately followed by the mask. Items in the brief condition were presented for 250 ms, and items in the long condition were presented for 2000 ms. The mask was presented for 2250 ms following a brief item, and was presented for 500 ms following a long item. Thus, the combined duration of the stimulus and the mask on each trial was 2500 ms. Immediately following the presentation of the mask, a central fixation (+) was presented for 500 ms[6]. As in Voss and Gonsalves (2010), participants were instructed to categorise each object as natural or manufactured by pressing one of two buttons. Speed was not emphasised for this response, and participants were told that they should respond on the basis of their initial

impression of the category, and that there was no correct answer. There were 170 study trials in total (85 brief, 85 long). The order of presentation of trials was randomized for each participant.

A retention interval of approximately 3 minutes followed the study phase, during which participants read instructions for the test phase. There were 320 trials in the test phase, comprising 85 brief items, 85 long items, and 150 new items. On each trial in the test phase an object was presented for 500 ms, followed by the same mask that was used in the study phase for 2000 ms. Participants were told that as soon as they saw the object at test they must categorize it as natural or manufactured by pressing one of two buttons, as in the study phase. As in Voss and Gonsalves (2010), speed was heavily emphasised for this response. Immediately after the mask was presented, an 'R' (for 'respond') was presented in the centre of the screen for 2000 ms. Participants were told that when this prompt was presented, they must make a second button press to indicate whether the object was presented before in the study phase (i.e., is "old") or was not presented before in the experiment (i.e., is "new"). They made this response by pressing one of four buttons corresponding to high confidence old, low confidence old, low confidence new, and high confidence new. It was emphasised to participants that for this response, accuracy was more important than speed. After the recognition prompt, a 500 ms break (a blank screen) followed before the next trial. Thus, as in Voss and Gonsalves (2010), but contrary to Experiments 1-3, test trials advanced automatically rather than being self-paced.

The order of presentation of trials in the test phase was randomized for each participant. The left and right index fingers were used for the natural and manufactured responses, and the assignment of responses to these fingers was counterbalanced across participants. The same assignment of classification responses to fingers was used at study and test for each participant. For one half of the participants, the high confidence new, low

confidence new, low confidence old, and high confidence old responses were assigned to the left middle, left index, right index, and right middle fingers respectively. For the other half of participants, the high confidence new, low confidence new, low confidence old, and high confidence old responses were assigned to the right middle, right index, left index and left middle fingers, respectively.

Results

Trials from the study phase were only analysed if a classification response had been registered (M = 96.50% of trials, SD = 8.27). Similarly, trials from the test phase were only analysed if both a classification response and a recognition judgment were registered (M = 88.22% of trials, SD = 18.02). Trials were also excluded from the analysis if the RT was greater than the mean RT plus two times the standard deviation RT, determined separately for study and test (study phase, M = 4.15% of trials excluded, SD = 1.69; test phase, M = 4.99% of trials excluded, SD = 1.05).

In contrast to Experiments 1-3 and Voss and Gonsalves (2010), the mean classification RT to brief, long, and new items did not significantly differ, $F(1.56, 60.97) = 0.66$, $p = .52$ (Table 1). Planned paired t-tests comparing brief and long items with new items indicated that there was no priming effect for either type of studied item (M brief = 7 ms, SE = 5; M long = 1 ms, SE = 7), $t(39) = 1.52$, $p = .14$, and $t(39) = 0.12$, $p = .91$, respectively, nor did RTs to brief and long items significantly differ, $t(39) = -0.80$, $p = .43$, $d_z = 0.13$ (the RT to brief items was shorter than the RT to long items in 23 out of 40 participants). The scaled JZS Bayes Factor for the comparison between classification RTs to brief and long items was 4.35 in favor of the null hypothesis, suggesting that the data are 4.35 times more likely to be observed under the null hypothesis (unplanned analysis).

The mean classification RT to brief and long items also did not significantly differ in the study phase, $t(39) = -0.87$, $p = .39$ (Table 3). Classification accuracy to brief, long and

new items did, however, significantly differ in the test phase (*M* brief = 93.83%, *SE* = 1.57; *M* long = 94.52%, *SE* = 1.35; *M* new = 92.55%, *SE* = 1.49; Table 5), $F(1.68, 65.69) = 3.72$, $p = .037$. This was unexpected, given our previous results in Experiments 1-3 and the results of Voss and Gonsalves (2010). Hence, (unplanned) paired *t*-tests with Bonferroni corrected *p*-values were conducted to compare accuracy to the three item types: these indicated that accuracy was significantly greater for long items than new items, $t(39) = 2.86$, $p = .020$, but that accuracy for new and brief items, although greater for brief items, did not significantly differ, $t(39) = 2.09$, $p = .13$. Accuracy for brief and long items did not differ, $t(39) = -0.79$, $p > .99$.

In contrast to priming, study duration affected recognition, as indicated by a significant ANOVA comparing the proportion of old judgments to brief, long, and new items, $F(1.22, 47.74) = 106.56$, $p < .001$ (Table 2). Planned *t*-tests indicated that the hit rate for brief and long items was significantly greater than the false alarm rate, $t(39) = 10.50$, $p < .001$, and $t(39) = 10.80$, $p < .001$, respectively, and the hit rate for long items was also greater than for brief items, $t(39) = 4.96$, $p < .001$, $d_z = 0.78$, indicating that recognition was better with a longer study duration (29 out of 40 participants showed this effect). Responses were analysed according to recognition confidence. A 3 (object: brief, long, new) x 4 (recognition response: high confidence old, low confidence old, low confidence new, high confidence new) within-subjects ANOVA yielded a significant object x recognition response interaction, $F(2.17, 84.57) = 70.09$, $p < .001$. As shown in Table 7, high confidence hits were more prevalent for long items than for brief items, $t(39) = 5.08$, $p < .001$. Low confidence hits were similar for long and brief items, $t(39) = -1.93$, $p = .061$. High confidence misses and low confidence misses were more prevalent for brief items than for long items, $t(39) = 3.12$, $p = .003$, and $t(39) = 3.11$, $p = .003$, respectively. This pattern of significance replicates Voss and Gonsalves (2010).

Recognition for fast and slow classification responses was also analysed, as in Voss and Gonsalves (2010). A within-subjects median split on classification RTs divided brief, long, and new trials into fastest and slowest categories. For brief, long and new item types, a 2 (condition: fastest, slowest) x 4 (response: high confidence old, low confidence old, low confidence new, high confidence new) repeated measures ANOVA was carried out on the mean proportion of recognition endorsements. As shown in Table 8, recognition accuracy tended to be better for slow responses than fast responses, and this was confirmed by the presence of a significant condition x response interaction for brief, $F(2.45, 95.34) = 7.85$, $p < .001$, and long, $F(2.47, 96.15) = 4.42$, $p = .009$, items. Voss and Gonsalves (2010) also found a significant interaction for brief and long items. Unlike Voss and Gonsalves (2010), the condition x response interaction for new items was not significant, $F(2.36, 92.18) = 0.83$, $p = .46$.

Classification RTs were also analysed according to the recognition judgment in order to test Predictions 2 and 3 of the model. Confirmation of these predictions was lacking, as was found in Experiments 1a, 2a, and 3a. There were no significant differences between classification RTs to items judged old or new within brief, long or new item types: $t(39) = -0.02$, $p = .99$, $t(39) = -1.99$, $p = .053$, and $t(39) = 0.15$, $p = .87$, respectively (Figure 4). The priming effect for items judged new was not significant for brief items ($M = 10$ ms, $SE = 8$), $t(39) = 1.26$, $p = .22$, but was significant for long items ($M = 16$ ms, $SE = 7$), $t(39) = 2.17$, $p = .036$. The priming effect did not significantly differ from the priming effect for items judged new for brief items, $t(39) = -0.37$, $p = .71$, or long items, $t(39) = -1.89$, $p = .067$.

Discussion

As found in Experiments 1-3 and also by Voss and Gonsalves (2010), recognition was greater following a long study duration than a short one. In contrast to the previous experiments and Voss and Gonsalves (2010), however, the priming effects for brief and long

items were not significant. There was also no significant difference in the mean classification RT for brief and long items. Thus, despite efforts to replicate Voss and Gonsalves (2010) as closely as possible, and having sufficient power, this experiment did not replicate their finding that study duration produces opposite effects on recognition and priming.

There were, if anything, indications of priming for long items and not brief items: classification accuracy was greater for long than new items, and the priming effect for items judged new was also reliable for long, but not brief, items. Nevertheless, one possibility is that the general lack of priming effects in this experiment is because our participants were less motivated than those in Voss and Gonsalves (2010). Indeed, recognition levels in Experiment 4 were much lower than in Voss and Gonsalves (2010) (hit minus false alarm rate, brief = 0.35, long = 0.43 in Experiment 4; vs. brief = 0.62, long = 0.73 in Voss and Gonsalves), and the effect of study duration on recognition was also smaller than reported by Voss and Gonsalves (2010) (d_z = 0.78 vs. d_z = 1.28) (though classification accuracy for brief, long and new items at test was slightly higher than in Voss and Gonsalves: 93.83%, 94.52%, 92.55%, respectively, vs. 91.2%, 89.9%, and 91.7%). Perhaps a similar pattern to Voss and Gonsalves (2010) would be observed in participants with recognition levels closer to those reported in their study? To investigate this we examined classification RTs in participants with the highest recognition scores for long items, as determined by a median split (the *M* long item hit minus false alarm rate of this sub-group = 0.62) (unplanned analysis). Small numerical priming effects for brief and long items were evident in this group, but the priming effect was, if anything, greater for long items as indicated by the shorter RTs for these items (*M* RT new = 751 ms, *M* RT brief = 747 ms, *M* RT long = 731 ms). Similarly, in the half of participants with the largest difference in hit rates to long and brief items (*M* long item hit rate minus *M* brief item hit rate = 0.16, d_z = 3.57), the priming effect was, if anything, greater for long items than brief items, again indicated by the shorter RTs for these items (*M* RT new

= 710 ms, *M* RT brief = 714 ms; *M* RT long = 709 ms) (unplanned analysis). Thus, there was no indication that the priming effect for brief items was greater than that of long items in participants whose recognition levels were more comparable to Voss and Gonsalves.

As in Voss and Gonsalves (2010), speed of responding was heavily emphasised at test for the classification response and the same RT trimming procedures were also followed. Despite this, RTs were still slower than in Voss and Gonsalves (2010) (*M* RT new items = 766 ms vs. *M* RT new items = 627 ms). RTs were, however, faster than in the earlier experiments that used the classification task (see Table 1), suggesting that these instructional and procedural changes did have some effect of speeding RTs. To investigate whether a similar pattern to Voss and Gonsalves (2010) would be observed in participants with more comparable RTs to those reported in their study, we looked at classification RTs in participants with the fastest classification RTs for new items, as determined by a median split (*M* RT new = 591 ms) (unplanned analysis). However, RTs were for long items were shorter than those of brief items in this group (*M* RT brief = 594 ms; *M* RT long = 592 ms).

Finally, a Bayesian analysis was also conducted in order to determine which outcome of the two studies—that of Experiment 4 or Voss and Gonsalves—should be considered more likely. For this, the Bayesian replication test described in Verhagen and Wagenmakers (2014) was used (unplanned analysis). Assuming that the methods used in Experiments 4 are adequate for the purposes of replicating Voss and Gonsalves' (2010) result, the Bayesian replication test assesses the likelihood of obtaining the data from Experiment 4 under the hypothesis that the brief-minus-long priming effect size is $d_z = 0.885$ (as found in Voss and Gonsalves) versus the hypothesis that the effect size is $d_z = 0.13$ (as found in Experiment 4). The results of this analysis are plotted in Figure 5. The curve with a dashed line represents the prior distribution (*M* = 0.855, *SD* = 0.31) for the effect of duration, given the result reported by Voss and Gonsalves. The curve with the solid black line represents the posterior

distribution (M = 0.28, SD = 0.14) for the effect of duration after the results of Experiment 4 have been taken into account. The gray and white dots indicate the ordinates of the prior and posterior distribution respectively at the hypothesis that the effect size is zero. The ratio of the heights of these two points gives the Bayes factor (denoted BF_{r0} = 0.07 in the figure), indicating that the data from Experiment 4 are 1/.07 = 14.3 times more likely under the null hypothesis of priming for long and brief items being equal, than under the alternative hypothesis of priming for brief items being greater than for long items. Thus, even if you believe the original Voss and Gonsalves (2010) effect then, from a Bayesian perspective, the new data should lead you to revise your belief so that the null hypothesis is now more likely than the Voss and Gonsalves hypothesis.

Modelling of Experiments 1-4

The data from Experiments 1 to 4 were modeled with the single-system model in order to determine how well it can simultaneously account for the main findings. As in previous applications of the model (e.g., Berry et al., 2012, 2014), numerous parameter values (from Equations 1 and 2) were fixed to enable identification of other (key) parameters: μ_{new}, the mean f of new items, was set to zero, as was $M(e_p)$, the mean of the noise associated with the priming task (Equation 2), and also $M(e_r)$, the mean of the noise associated with the recognition task (Equation 1). σ_f, the standard deviation of f, was set to equal $\sqrt{0.5}$; σ_r, the standard deviation of e_r, was set to equal σ_f (and so $\sigma_{Jr} = \sqrt{(\sigma_f^2 + \sigma_f^2)} = 1$, which means that μ for an old item condition is therefore equal to d' for that condition). For Experiments 1a, 1b, 2a, and 2b, the model has 6 free parameters: μ_{250ms}, the mean f of items in the 250 ms condition at test; μ_{2250ms}, the mean f value of items in the 2250 ms condition; C, the criterion of J_r that is required for an "old" judgment to be made; b, s, and σ_p, the standard deviation of e_p, which is the noise added to f for the generation of the RT (in Equation 2). For Experiments

3a and 3b, there were 10 free parameters in total: *C*, *b*, *s*, and σ_p, in addition to the parameters μ_{40ms}, μ_{150ms}, μ_{250ms}, μ_{350ms}, μ_{500ms}, μ_{1900ms}, which represent the mean *f* of the respective old item condition. For Experiment 4, there were 6 free parameters, as Experiments 1 and 2, μ_{250ms}, μ_{2000ms}, *C*, *b*, *s*, and σ_p.

The parameters of the model were estimated using maximum likelihood estimation. This involved determining the likelihood of each recognition judgment ("old" or "new") and RT of each item at test, given particular parameter values of the model. The likelihood of each judgment (*Z*) and response time (*RT*) combination at test is given as:

$$L(Z, RT \mid X) = [\Phi(C_j \mid \mu_{Jr|RT,X}, \sigma^2_{Jr|RT}) - \Phi(C_{j-1} \mid \mu_{Jr|RT,X}, \sigma^2_{Jr|RT})]\phi(RT \mid b - s\mu_X, \sigma^2_{RT})$$

where *X* = new, 250 ms, 2250 ms for Experiments 1 and 2, *X* = new, 40 ms, 150 ms, 250 ms, 350 ms, 500 ms, 1900 ms for Experiment 3, and *X* = new, 250 ms, 2000 ms for Experiment 4; Φ is the cumulative normal distribution function, $j = 1$ when $Z =$ "new", $j = 2$ when $Z =$ "old"; $C_0 = -\infty$, $C_1 = C$, $C_2 = \infty$; ϕ is the normal density function; $\sigma_{RT}^2 = s^2\sigma_f^2 + \sigma_p^2$ (from Equation 2); and $\mu_{Jr|RT}$ and $\sigma^2_{Jr|RT}$ are the mean and variance of the conditional distribution of J_r given *RT*, where:

$$\mu_{Jr|RT,X} = \mu_X - \frac{s\sigma_f^2\left(RT - b + s\mu_X\right)}{s^2\sigma_f^2 + \sigma_p^2},$$

and

$$\sigma^2_{Jr|RT} = \sigma_f^2 + \sigma_r^2 - \frac{s^2\sigma_f^4}{s^2\sigma_f^2 + \sigma_p^2}$$

The parameter values that maximised the summed log-likelihood across all test phase trials were determined via an automated search process (goodness-of-fit values are shown in Table 9). For each experiment, the same participants that were used in the analysis were also modelled, however, parameters could not be estimated if a participant had zero responses in any of the hit, miss, false alarm and correct rejection categories, which meant that there was 1 participant in Experiment 1a (N = 31), one participant from Experiment 2a (N = 31), and 3 participants from Experiment 3b (N = 26) that could not be fit by the model. Of the remaining participants, the same trials that were analysed to produce the findings in the results sections were also the trials that were fit by the model (i.e., trials on which the RT was an outlier were excluded, and incorrectly identified CID-R trials were also excluded). In total, 5862 trials were fit in Experiment 1a, 4470 trials in Experiment 1b, 5851 trials in Experiment 2a, 3207 trials in Experiment 2b, 11956 trials in Experiment 3a, 5139 trials in Experiment 3b, and 10732 trials in Experiment 4. One set of parameters was estimated for each individual participant, and the means of the maximum likelihood estimates of the parameters across individuals are given in Table 10.

Certain trends in the parameter estimates are worth noting: First, estimates of μ tended to increase with study duration. One exception to this was in Experiment 3b where the estimate of μ_{350ms} was lower than the estimate of μ_{250ms}; this reflected the trend for both priming and recognition in the 250 ms condition to be slightly greater than in the 350 ms condition. The value of the response criterion C was fairly consistent across experiments. The values of b, s and σ_p tended to be greater for the experiments using the CID-R task than the classification task, which reflected the generally longer RTs and greater variance in RTs in this task (e.g., Table 1, Equation 2).

The estimates of the model parameters were used to calculate expected results of the single-system model. Expected priming was calculated for each condition as the difference in

the expected RT for new and old items, that is, the expected priming effect = $b - (b - s\mu_X) = s\mu_X$, where X denotes the old item condition (e.g., 250 ms); the hit rate is calculated as $1 - \Phi(C - \mu_X)$. The expected RT for hit, miss, false alarm and correct rejection responses is calculated with the formula:

$$E[RT|Z,\ X] = b - s\mu_X + \frac{s\sigma_f^2}{\sigma_{Jr}} \frac{\phi\left(\frac{C_j - \mu_X}{\sigma_{Jr}}\right) - \phi\left(\frac{C_{j-1} - \mu_X}{\sigma_{Jr}}\right)}{\Phi\left(\frac{C_j - \mu_X}{\sigma_{Jr}}\right) - \Phi\left(\frac{C_{j-1} - \mu_X}{\sigma_{Jr}}\right)}$$

where $\sigma_{Jr}^2 = \sigma_f^2 + \sigma_r^2$, and, again, X = new, 250 ms, 2250 ms for Experiments 1 and 2, X = new, 40 ms, 150 ms, 250 ms, 350 ms, 500 ms, 1900 ms for Experiment 3, and X = new, 250 ms, 2000 ms for Experiment 4; j = 1 when Z = "new", j = 2 when Z = "old"; $C_0 = -\infty$, $C_1 = C$, $C_2 = \infty$. The formula gives the expected RT for false alarms when Z = "old" and X = new, and the expected RT for correct rejections when Z = "new" and X = new. The expected RT for hits and misses is then given when Z = "old" and Z = "new", respectively, and X is the relevant old item condition (e.g., 250 ms). The difference in the expected RT for hits and misses and the difference between the expected RT for correct rejections and false alarms can then be calculated, and the mean expected difference is plotted alongside the data in Figures 2 and 4. Also plotted in Figures 2 and 4 is the expected difference in the magnitude of the priming effect and the priming effect for items judged new, which can be calculated for each old item condition as $s\mu_X$ – (E[RT| Z = "new", X = new] − E[RT| Z = "new", X = old]). Expected values were calculated for each individual, and the mean of the expected values across participants is plotted as open circles in Figures 1 to 4.

The model successfully reproduced the major patterns in the data: the predicted effects of study duration were qualitatively similar for recognition and priming (Figures 1 and 3). Pooling the model data from the 250 ms conditions and longest duration conditions from

Experiments 1-4, the magnitude of the predicted effect on priming also tended to be smaller than the effect on recognition (classification task: priming $d_z = 0.46$ vs. recognition $d_z = 1.17$; CID-R task: priming $d_z = 0.53$ vs. recognition $d_z = 0.60$). The model also reproduced the smaller mean difference in priming between brief and long duration conditions for the experiments with the classification task than the CID-R task (e.g., Figures 1a and 1c). It does this because estimates of *s* are lower for the classification task (Table 10), and the value of *s* directly affects the magnitude of the expected priming effect, $s\mu_X$, and also the expected difference in priming between duration conditions, $s(\mu_{long} - \mu_{brief})$. (Note that *s* also affects the expected variance of RT, $\sigma_{RT}^2 = s^2\sigma f^2 + \sigma_p^2$.)

Considering Experiment 4 specifically, the mean estimates of *s*, σ_p, and *b* were all lower than in Experiments 1a, 2a, and 3a, consistent with the generally shorter RTs in this experiment using the classification task. The difference between μ in the brief and long conditions was also smaller than in Experiments 1a, 2a, and 3a. In other words, according to the model, the effect of study duration is smallest in this experiment, and this translates into a small predicted effect on priming ($d_z = 0.07$) and recognition ($d_z = 0.86$).

Also, where Predictions 2 and 3 were not upheld in the data, the model generally predicted small differences in RTs (Figures 2a, c, e; Figure 4), whereas in cases where they were upheld, the model predicted larger differences (Figures 2b, d, f). Again, the reason the model predicted this is because estimates of *s* are smaller for the classification task than the CID-R task. A larger value of *s* produces greater differences in RTs between responses with different expected values of *f*, such as hits and misses (Equation 2). It is worth noting here that the value of *s* for the classification task is not only small relative to estimates of *s* for the CID-R task, but also small relative to the estimates of σ_p and *b* for the classification task (i.e., the other parameters used to generate RTs). One way of interpreting this from the model's perspective is that the sensitivity of RT to *f* in the classification task is low relative to the

CID-R task. Thus the model proposes that, with respect to measuring priming, a key difference between the two experimental tasks employed here is that they differ substantially in the extent to which a given change in memory strength *f* translates into a change in response time.

Voss and Gonsalves (2010) provided further support for an inverse relationship between priming and recognition by showing that recognition accuracy was poorer for items with relatively fast priming task RTs than those with relatively slow priming task RTs (where the fast/slow categories were defined by a median split). That is, the recognition hit rate in the brief and long conditions was greater for items that were identified slowly (showed less priming), and the false alarm rate to new items was lower when items were identified slowly (Voss & Gonsalves, 2010; Table 2). In contrast the model predicts that an item with a fast RT is likely to have a relatively high value of *f* (Equation 2) and therefore is also likely to have a high value of J_r and be judged old. Thus, it predicts that items identified quickly are also likely to be judged old, regardless of the item type (brief, long, or new) (similar to Prediction 2).

In the results of Experiment 4, we reported the results from the same analysis, and found the same pattern of results reported by Voss and Gonsalves (2010) (for brief and long items, but not new items). This is somewhat inconsistent with the single-system model predictions, though it is not completely clear what to make of these results, given that there were no significant differences in RTs for brief, long, and new RTs in this particular experiment. The data from Experiments 1 to 3 were pooled and also analysed in the same way as in Voss and Gonsalves and the results are presented in Table 11. For the classification task (Experiments 1a, 2a, and 3a pooled), there was no significant difference in the tendency to endorse brief or long items as old between items with fast priming task RTs versus slow priming task RTs (brief, $t(92) = 0.71$, $p = .48$, 95% CI [-.02, .04]; long, $t(92) = -0.80$, $p = .43$,

95% CI [-.04, .02]). The false alarm rate was, however, significantly greater for items identified quickly, $t(92) = 2.03$, $p = .045$, 95% CI [.0005, .04], as reported by Voss and Gonsalves, and as predicted by the model. For the CID-R task (Experiments 1b, 2b, and 3b pooled), the tendency to judge an item as old was significantly greater if the item was identified quickly than slowly, and this was the case for brief, $t(70) = 4.65$, $p < .001$, 95% CI [.05, .12], long, $t(70) = 2.85$, $p = .006$, 95% CI [.02, .09], and new, $t(70) = 3.86$, $p < .001$, 95% CI [.03, .10], items. Thus, there was no evidence for an inverse relationship between priming and recognition in Experiments 1-3. Instead, faster RTs tended to lead to a greater likelihood of an old judgment in the CID-R task, in line with the model predictions, but there was no evidence of this in the classification task.

General Discussion

Across seven experiments, we found no evidence that recognition memory and repetition priming are affected in opposite ways by varying the duration with which items are initially studied. Instead, we found that, although longer study durations improved subsequent recognition, they either had no detectable effect on priming (Experiments 1a, 2a, 4), or also increased priming (Experiment 1b, 2b, 3a, 3b). The effect of study duration on priming in the CID-R task was particularly robust, being detected in all experiments using this task (Experiments 1b, 2b, and 3b), whereas the effect on priming in the classification task was only found in Experiment 3a. Where an effect of study duration on priming was found, it tended to be smaller than on recognition, consistent with the effects that we have found other variables to have on recognition and priming, such as normal aging (Ward et al., 2013a, 2013b), selective attention at encoding (Berry et al., 2006), and amnesia (Berry et al., 2014).

The single-system model predicts that study duration will have qualitatively similar effects on recognition and priming (Prediction 1, see the *Introduction*). The model assumes that the magnitudes of recognition and priming are positive functions of μ, the mean strength

of the underlying memory signal. Since maximum likelihood estimates of μ tended to be greater for longer study duration conditions (Table 10), the model predicted that both recognition and priming would increase with study duration. When fit to the data, the model reproduced the smaller effect on priming by assuming that there is a greater influence of noise in the measurement of the priming effect, compared to recognition. That is, the influence of e_p (in Equation 2) is greater than the influence of e_r (in Equation 1) because σ_p is larger relative to σ_f than σ_r is relative to σ_f. The magnitude of the effect of experimental variables on priming will therefore tend to be smaller than on recognition (see also Meier & Perrig, 2000; Buchner & Wippich, 2000, for a similar account).

We also evaluated two more subtle sets of predictions made by the single-system model. These are important because they capture predicted associations, rather than dissociations, between priming and recognition at the item level. For example, the model predicts that RTs will be faster for hits than misses, and faster for false alarms than correct rejections (as can be seen in Figure 2). Since hits and false alarms represent recognition (i.e., "old") responses, and faster RTs reflect greater priming, these predictions entail positive associations between priming and recognition in contrast to the idea that these manifestations of memory are independent or indeed (as claimed by Voss & Gonsalves) negatively related. Support for Predictions 2 and 3 of the single-system model (see *Introduction*) was obtained, but only with the CID-R task. In support of Prediction 2, identification RTs in the CID-R task to items judged old tended to be significantly shorter than those of items judged new, regardless of whether the item had been presented for a brief or long duration in the study phase, or whether the item was new. In support of Prediction 3, within brief and long item conditions, the priming effect also tended to be greater than the priming effect for items judged new. Confirmation of these predictions has been important in providing evidence against a multiple-systems version of the model in which the memory signals driving

recognition and priming are completely independent and indicate that, if there are indeed multiple systems, then there is a high degree of association between them (see Berry et al., 2012; 2014; Ward et al., 2013a). The results regarding Predictions 2 and 3 were not significant in the classification task. However, trends in the direction predicted by the single-system model were evident in many cases (e,g., the majority of bars in Figures 2c and 2e are greater than zero) and given that the effects predicted by the model were small relative to the CID-R task it is difficult to exclude the possibility that a failure to detect these effects is due to low power. Indeed, the power of Experiments 1a, 2a, 3a, and 4 (individually) to detect a small effect of Cohen's $d = 0.2$ was 0.23 or lower (two-tailed test).

In our study priming was found to be a monotonically increasing function of study duration, but might we expect priming to be a nonmonotonic function of study duration on the basis of other research? For example, as mentioned previously, Zago et al. (2005) found the magnitude of priming to rise and then fall across study durations of 40 ms to 1900 ms, peaking at 250 ms. Miyoshi and Ashida (2014) also found greater priming for items presented for 250 ms compared to 350 ms at study. In the Zago et al. (2005) study, however, recognition was not measured and so it is unknown whether recognition would show the same relationship with study duration when the same procedures are used. Miyoshi and Ashida (2014) did measure recognition memory, but did not find that recognition followed a different function. Their focus was on recognition accuracy for guesses (as this may provide a measure of implicit recognition), which was greater for items presented for 250 ms than 350 ms, and so showed the same pattern as priming. It is important to emphasise here that the single-system model does not necessarily predict that the magnitude of recognition and priming are monotonically increasing functions of study duration; the model only predicts that the function relating recognition and priming is monotonically increasing. That is, it could be that under some specific experimental conditions priming is a nonmonotonic

function of study duration, but the important point is that the model predicts that recognition would also be a nonmonotonic function of study duration under the same conditions. To provide strong evidence against the model it would be necessary to provide evidence that the function relating the magnitude of recognition and priming is nonmonotonic. Robust evidence of a reversed association could potentially be sufficient for this purpose (Dunn & Kirsner, 1988).

Finally, in this article we have not considered a large body of neuroscientific and neuropsychological research, which is often taken to favour the view that recognition and priming—and explicit and implicit memory more broadly—have distinct neural correlates and are driven by distinct brain regions (for a review see e.g., Reber, 2013). A full consideration of this work is beyond the scope of this article. Our focus has been on testing the single-system model, which makes behavioural predictions and has not been extended to neural data. We make two points, however. First, the present research shows that a single-system model is able to generate a pattern of findings in which an independent variable has a much larger effect on recognition than on priming. This is precisely the famous pattern seen in individuals with amnesia (Schacter et al., 1993). Elsewhere (Berry et al., 2012, 2014), we have fitted the model to data from individuals with amnesia and shown that it is readily able to simulate the canonical patterns of priming and recognition that are observed.

Secondly, recent research has suggested that the neural distinction between explicit and implicit memory may not be as sharp as is widely believed, with numerous imaging studies reporting overlap in the brain regions associated with item recognition and repetition priming (Gomes, Figueiredo, & Mayes, 2016; Thakral, Kensinger, & Slotnick, 2015; Turk-Browne, Yi, & Chun, 2006), and a variety of findings suggesting that the medial temporal lobes—long thought only to be important for declarative memory—are actually important for implicit/nondeclarative memory too (e.g., Addante, 2015; Berry et al., 2014; Hannula &

Greene, 2012; Henke, 2010). Recognition and priming may rely on a common underlying memory representation that is distributed across multiple brain regions, but this representation may be accessed differently according to the demands of each task. The unidimensional strength signal in the single-system model can be conceptualised as characterising the influence of this memory representation on performance in priming and recognition tasks.

To conclude, we found that studying an item for a long versus a brief duration at encoding either 1) improves both recognition and priming, with the effect being smaller for priming, or 2) improves recognition, but not priming. Thus, we find no evidence that longer durations have opposite effects on recognition and priming, as found by Voss and Gonsalves (2010). A high-powered direct replication of Voss and Gonsalves (2010) in Experiment 4 also found that a longer duration improves recognition, but not priming (although it is important to note that priming effects for brief and long items were not reliable in this experiment overall). Presupposing that the greater priming for long versus brief items in Voss and Gonsalves (2010) is not a Type I error, it is possible that some outstanding methodological difference in our experiments, and Experiment 4 in particular, led to the failure to replicate (e.g., participant motivation, stimulus set). The identification of the precise conditions that give rise to the effect therefore remains a goal for future research. Until such conditions are identified, however, we regard our findings as being consistent with the single-system model, which was able to reproduce the main trends in the data and also provides a parsimonious account in that it assumes that recognition and priming are driven by the same memory strength signal or system, rather than distinct explicit and implicit ones.

Acknowledgements

The data for each experiment are available on the Open Science Framework at https://osf.io/7xzas/. This work was supported by the Economic and Social Research Council (ES/N009916/1).

References

Addante, R. J. (2015). A critical role of the human hippocampus in an electrophysiological measure of implicit memory. *NeuroImage, 109,* 515-528.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Caski (Eds.), *Second international symposium on information theory,* pp. 267–281. Budapest: Academiai Kiado.

Berry, C. J., Henson, R. N. A., & Shanks, D. R. (2006). On the relationship between repetition priming and recognition memory: Insights from a computational model. *Journal of Memory and Language, 55,* 515–533.

Berry, C. J., Shanks, D. R., & Henson, R. N. A. (2008). A single-system account of the relationship between priming, recognition, and fluency. *Journal of Experimental Psychology*: *Learning, Memory, and Cognition, 34,* 97–111.

Berry, C. J., Kessels, R. P., Wester, A. J., & Shanks, D. R. (2014). A single-system model predicts recognition memory and repetition priming in amnesia. *The Journal of Neuroscience*, *34*(33), 10963-10974.

Berry, C. J., Shanks, D. R., Speekenbrink, M., & Henson, R. N. A. (2012). Models of recognition, repetition priming, and fluency: Exploring a new framework. *Psychological Review, 119,* 40–79.

Blaxton, T. A. (1989). Investigating dissociations among memory measures: Support for a transfer-appropriate processing framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 657-668.

Brodeur, M. B., Guérard, K., & Bouras, M. (2014). Bank of standardized stimuli (BOSS) phase II: 930 new normative photos. *PloS One*, *9*(9), e106953.

Brown, A. S., & Mitchell, D. B. (1994). A reevaluation of semantic versus nonsemantic processing in implicit memory. *Memory & Cognition*, *22*(5), 533-541.

Buchner, A., & Wippich, W. (2000). On the reliability of implicit and explicit memory measures. *Cognitive Psychology, 40*(3), 227-259.

Craik, F. I. M., Moscovitch, M., & McDowd, J. M. (1994). Contributions of surface and conceptual information to performance on implicit and explicit memory tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(4), 864-875.

Dew, I. T., & Cabeza, R. (2011). The porous boundaries between explicit and implicit memory: behavioral and neural evidence. *Annals of the New York Academy of Sciences, 1224*(1), 174-190.

Dew, I. T., & Mulligan, N. W. (2008). The effects of generation on auditory implicit memory. *Memory & Cognition, 36*(6), 1157-1167.

Dunn, J. C. (2003). The elusive dissociation. *Cortex, 39*, 177–179.

Dunn, J. C., & Kirsner, K. (1988). Discovering functionally independent mental processes: The principle of reversed association. *Psychological Review, 95*(1), 91-101.

Eakin, D. K., & Smith, R. (2012). Retroactive interference effects in implicit memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(5), 1419-1424.

Feustel, T. C., Shiffrin, R. M., & Salasoo, A. (1983). Episodic and lexical contributions to the repetition effect in word identification. *Journal of Experimental Psychology: General, 112*(3), 309-346.

Gabrieli, J. D. E. (1998). Cognitive neuroscience of human memory. *Annual Review of Psychology, 49*, 87-115.

Gomes, C. A., Figueiredo, P., & Mayes, A. (2016). Priming for novel object associations: Neural differences from object item priming and equivalent forms of recognition. *Hippocampus, 26*, 472-491.

Hamann, S. B., & Squire, L. R. (1997). Intact perceptual memory in the absence of conscious memory. *Behavioral Neuroscience, 111,* 850–854.

Hannula, D. E., & Greene, A. J. (2012). The hippocampus reevaluated in unconscious learning and memory: At a tipping point? *Frontiers in Human Neuroscience, 6*.

Henke, K. (2010). A model for memory systems based on processing modes rather than consciousness. *Nature Reviews Neuroscience*, *11*(7), 523-532.

Horner, A. J., & Henson, R. N. (2009). Bindings between stimuli and multiple response codes dominate long-lag repetition priming in speeded classification tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 757-779.

Henson, R. N., Eckstein, D., Waszak, F., Frings, C., & Horner, A. J. (2014). Stimulus–response bindings in priming. *Trends in Cognitive Sciences*, *18*(7), 376-384.

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

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General*, *110*(3), 306-340.

MacLeod, C. M. (2008). Implicit memory tests: Techniques for reducing conscious intrusion. *Handbook of metamemory and memory*. (pp. 245-263) Psychology Press, New York, NY.

Macmillan, N. A., & Creelman, C. D. (2005*). Detection theory: A user's guide (2nd ed.)* Lawrence Erlbaum Associates Publishers, Mahwah, NJ.

Masson, M. E. J., & MacLeod, C. M. (1992). Reenacting the route to interpretation: Enhanced perceptual identification without prior perception. *Journal of Experimental Psychology: General*, *121*, 145-176.

Meier, B., & Perrig, W. J. (2000). Low reliability of perceptual priming: Consequences for the interpretation of functional dissociations between explicit and implicit memory. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *53*A(1), 211-233.

Miyoshi, K., & Ashida, H. (2014). Priming and implicit recognition depend on similar temporal changes in perceptual representations. *Acta Psychologica*, *148*, 6-11.

Miyoshi, K., Kimura, Y., & Ashida, H. (2015). Longer prime presentation decreases picture–word cross-domain priming. *Frontiers in Psychology*, *6*.

Mulligan, N. W., & Dew, I. T. Z. (2009). Generation and perceptual implicit memory: Different generation tasks produce different effects on perceptual priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(6), 1522-1538.

Mulligan, N. W., & Osborn, K. (2009). The modality-match effect in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(2), 564-571.

Musen, G. (1991). Effects of verbal labeling and exposure duration on implicit memory for visual patterns. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(5), 954-962.

Neill, W. T., Beck, J. L., Bottalico, K. S., & Molloy, R. D. (1990). Effects of intentional versus incidental learning on explicit and implicit tests of memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(3), 457-463.

Ostergaard, A. L. (1998). The effects on priming of word frequency, number of repetitions, and delay depend on the magnitude of priming. *Memory & Cognition*, *26*(1), 40-60.

Ostergaard, A. L. (1999). Priming deficits in amnesia: now you see them, now you don't. *Journal of the International Neuropsychological Society*, *5*, 175-190.

Reber, P. J. (2013). The neural basis of implicit learning and memory: A review of neuropsychological and neuroimaging research. *Neuropsychologia, 51*, 2026-2042.

Reder, L. M., Park, H., & Kieffaber, P. D. (2009). Memory systems do not divide on consciousness: Reinterpreting memory in terms of activation and binding. *Psychological Bulletin, 135*(1), 23-49.

Ryan, J. D., & Cohen, N. J. (2003). Evaluating the neuropsychological dissociation evidence for multiple memory systems. *Cognitive, Affective & Behavioral Neuroscience, 3*(3), 168-185.

Schacter, D. L., Chiu, C. -. P., & Ochsner, K. N. (1993). Implicit memory: A selective review. *Annual Review of Neuroscience, 16*, 159-182.

Schwarz, G. (1978). Estimating the dimension of a model, *The Annals of Statistics, 6*, 461–464.

Shanks, D. R., & Berry, C. J. (2012). Are there multiple memory systems? Tests of models of implicit and explicit memory. *The Quarterly Journal of Experimental Psychology, 65*(8), 1449-1474.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26*(5), 559-569.

Squire, L. R. (2004). Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning and Memory, 82*(3), 171-177.

Squire, L. R., & Dede, A. J. (2015). Conscious and Unconscious Memory Systems. *Cold Spring Harbor Perspectives in Biology, 7*, a021667.

Stark, C. E. L., & McClelland, J. L. (2000). Repetition priming of words, pseudowords, and nonwords. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 945–972.

Thakral, P. P., Kensinger, E. A., & Slotnick, S. D. (2016). Familiarity and priming are mediated by overlapping neural substrates. *Brain Research, 1632,* 107-118.

Tulving, E., & Schacter, D. L. (1990). Priming and human memory systems. *Science, 247*(4940), 301-306.

Turk-Browne, N. B., Yi, D. J., & Chun, M. M. (2006). Linking implicit and explicit memory: common encoding factors and shared representations. *Neuron, 49,* 917-927.

Verhagen, J., & Wagenmakers, E. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General, 143,* 1457-1475.

von Hippel, W., & Hawkins, C. (1994). Stimulus exposure time and perceptual memory. *Perception & Psychophysics, 56*(5), 525-535.

Voss, J. L., & Gonsalves, B. D. (2010). Time to go our separate ways: opposite effects of study duration on priming and recognition reveal distinct neural substrates. *Frontiers in Human Neuroscience, 4,* 1–11.

Ward, E. V., Berry, C. J., & Shanks, D. R. (2013a). Age effects on explicit and implicit memory. *Frontiers in Psychology, 4*.

Ward, E. V., Berry, C. J., & Shanks, D. R. (2013b). An effect of age on implicit memory that is not due to explicit contamination: Implications for single and multiple-systems theories. *Psychology and Aging, 28*(2), 429-442.

Wichmann, F. A., Sharpe, L. T., & Gegenfurtner, K. R. (2002). The contributions of color to recognition memory for natural scenes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(3), 509-520.

Zago, L., Fenske, M.J., Aminoff, E., & Bar, M. (2005). The rise and fall of priming: how visual exposure shapes cortical representations of objects. *Cerebral Cortex, 15,* 1655–665.

Footnotes

[1] The analyses reported in the Results sections of Experiment 1-3 were repeated except with items at test removed from the analyses if they were categorized incorrectly either at study or test. We did not do this for Experiment 4, where we instead followed the same analysis method as Voss and Gonsalves (2010). For this analysis, it is worth bearing in mind that there is a degree of subjectivity in the classification judgment and also that the removal of trials results in decreased statistical power. The proportion of each trial type that was removed because of an incorrect classification response at test (in Experiments 1a, 2a, and 3a) is given in Table 5. Of the remaining trials, the percentage of old trials at test that were excluded because the item on those trials was incorrectly classified at study was as follows: Experiment 1a, M = 10.01%, SE = 2.19; Experiment 1b, M = 7.56%, SE = 1.22; Experiment 2a, M = 5.50%, SE = 0.56%; Experiment 2b, M = 6.89%, SE = 1.22; Experiment 3a, M = 16.13%, SE = 1.55%; Experiment 3b, M = 21.21%, SE = 1.78. Minor differences in the remaining experiments are reported below.

The results are presented here for comparison, but the conclusions drawn in the main article are unaffected. In Experiment 1a, the only difference in results was that RTs to misses in the 2250 ms condition were now significantly faster than those of correct rejections, $t(29) = 2.07$, $p = .048$. In Experiment 1b, the difference in priming between the conditions was no longer significant, $t(23) = 1.64$, $p = .11$, although priming was still numerically greater in the 2250 ms condition. There was no difference in the pattern of significance in Experiments 2a or 2b. In Experiment 3a, the repeated measures ANOVA comparing the magnitude of priming across the 40 ms, 150 ms, 250 ms, 350 ms, 500 ms, and 1900 ms conditions was no longer significant, $F(3.44, 96.17) = 1.55$, $p = .18$, although the linear trend across these condition remained significant, $F(1,28) = 7.31$, $p = .012$. Also, while numerically greater, the magnitude of priming in the 1900 ms condition was no longer significantly greater than in the

250 ms condition, $t(28) = 1.61$, $p = .12$. Finally, the difference between the RTs to correct rejections and misses to items in the 40 ms condition was now significant, $t(28) = 2.43$, $p = .022$. In Experiment 3b, when RTs were broken down according to the recognition judgments, the priming effect for items judged new was only marginally significant in the 250 ms, $t(23) = 2.05$, $p = .052$, and 1900 ms, $t(23) = 2.06$, $p = .051$, conditions.

[2] Voss and Gonsalves (2010) did not report what proportion of study items were not classified within the study trial duration.

[3] Identification RTs tended to be longer in Experiment 3b than Experiments 1b and 2b (Table 1). This is likely due to 1) the size of the stimuli being smaller, which may have made identification more difficult, and 2) the longer duration of the experiment, arising from the greater number of trials in the study and test phases, which may have impacted upon levels of motivation/fatigue in participants.

[4] Though it is worth bearing in mind that items with RTs less than 200 ms or greater than three standard deviations from a participant's mean RT were not included in the analysis of Experiments 1-3 (as in Berry et al., 2012; see *Methods*). Voss and Gonsalves (2010) screened RTs by calculating the RT mean and *SD* for each participant and then removing individual trials if the RT was greater than two standard deviations from the mean RT for the participant collapsed across all conditions (separately for study and test) (J. L. Voss, personal communication, 2016). These differences in the screening of RTs between studies, and also the fact that trials at test only advanced once a response was made, rather than automatically (as in Voss and Gonsalves, 2010), may of course account for the slightly longer baseline RTs in Experiments 1-3.

[5] The images in our stimulus pool were selected so that they were relatively unambiguous with regards to whether they depicted a natural or manufactured object. We wanted to control for this as much as possible because in Voss and Gonsalves' (2010) study a

classification response to an item was scored as correct if it matched the modal response for that item across participants, which implies a degree of ambiguity in the objective classification of their stimuli. Only one stimulus in Experiment 4 differed in terms of its objective and modal classification at test.

[6] Although the presentation of this fixation is not mentioned in the Materials and Methods section of Voss and Gonsalves (2010), a brief fixation separating trials was used in their study phase (B. Gonsalves, personal communication, 2016).

Table 1.

Mean RTs (ms) in the Priming Task in the Test Phase

	Study duration (ms)								
Experiment and task	New	40	150	250	350	500	1900	2000	2250
1a									
Classification	893 (45)	-	-	859 (43)	-	-	-	-	859 (42)
1b									
CID-R	2426 (92)	-	-	2169 (96)	-	-	-	-	2074 (95)
2a									
Classification	986 (63)	-	-	943 (61)	-	-	-	-	922 (54)
2b									
CID-R	2437 (116)	-	-	2173 (113)	-	-	-	-	2038 (86)
3a									
Classification	1057 (56)	1040 (54)	1018 (49)	1021 (53)	1005 (53)	985 (47)	985 (48)	-	-
3b									
CID-R	3306 (114)	3270 (112)	3004 (110)	2921 (109)	2993 (93)	2941 (108)	2875 (91)	-	-
4									
Classification	766 (33)	-	-	759 (33)	-	-	-	766 (33)	-

Note. Standard errors are in parentheses. A dash indicates that this condition was not included in the experiment.

Table 2.

Mean Proportion of "Old" Recognition Judgments in the Test Phase

	Study duration (ms)								
Experiment and task	New	40	150	250	350	500	1900	2000	2250
1a									
Classification	.17 (.02)	-	-	.78 (.02)	-	-	-	-	.88 (.01)
1b									
CID-R	.20 (.02)	-	-	.71 (.03)	-	-	-	-	.81 (.02)
2a									
Classification	.15 (.01)	-	-	.73 (.02)	-	-	-	-	.85 (.02)
2b									
CID-R	.23 (.04)	-	-	.67 (.04)	-	-	-	-	.73 (.04)
3a									
Classification	.28 (.03)	.33 (.03)	.58 (.02)	.65 (.02)	.69 (.02)	.71 (.02)	.75 (.02)	-	-
3b									
CID-R	.27 (.04)	.28 (.03)	.54 (.04)	.58 (.04)	.57 (.03)	.60 (.05)	.63 (.05)	-	-
4									
Classification	.28 (.03)	-	-	.63 (.02)	-	-	-	.71 (.02)	-

Note. Standard errors are in parentheses. A dash indicates that this condition was not included in the experiment.

Table 3.

Mean Classification RTs (ms) in the Study Phase

	Study duration (ms)							
Experiment	40	150	250	350	500	1900	2000	2250
1a	-	-	691 (22)	-	-	-	-	682 (26)
1b	-	-	708 (20)	-	-	-	-	705 (17)
2a	-	-	733 (35)	-	-	-	-	761 (47)
2b	-	-	955 (67)	-	-	-	-	991 (83)
3a	698 (27)	647 (17)	648 (18)	639 (16)	638 (19)	640 (18)	-	-
3b	704 (27)	653 (19)	669 (17)	661 (18)	649 (19)	678 (27)	-	-
4	-	-	715 (21)	-	-	-	749 (45)	-

Note. Standard errors are in parentheses. A dash indicates that this condition was not included in the experiment.

Table 4.

Mean Percentage of Classification Errors in the Study Phase

	Study duration (ms)						
Experiment	40	150	250	350	500	1900	2250
1a	-	-	8.09 (1.67)	-	-	-	8.26 (2.28)
1b	-	-	5.35 (1.14)	-	-	-	7.41 (1.22)
2a	-	-	4.80 (0.69)	-	-	-	5.22 (0.61)
2b	-	-	7.11 (1.34)	-	-	-	5.55 (1.34)
3a	23.51 (2.27)	9.88 (1.92)	9.49 (1.40)	9.10 (1.72)	8.73 (1.42)	8.24 (1.08)	-
3b	28.63 (2.64)	10.46 (1.35)	9.17 (1.28)	11.75 (1.51)	7.70 (1.18)	11.92 (1.97)	-

Note. Standard errors are in parentheses. A dash indicates that this condition was not included in the experiment.

Table 5.

Mean Error Rate (%) in the Test Phase

Experiment and task	Study duration (ms)								
	New	40	150	250	350	500	1900	2000	2250
1a									
Classification	5.79 (0.94)	-	-	6.09 (1.19)	-	-	-	-	5.55 (0.91)
1b									
CID-R	1.20 (0.49)	-	-	1.04 (0.79)	-	-	-	-	1.32 (0.42)
2a									
Classification	8.28 (0.91)	-	-	5.22 (0.65)	-	-	-	-	6.93 (0.94)
2b									
CID-R	4.41 (0.92)	-	-	4.61 (1.18)	-	-	-	-	4.62 (1.06)
3a									
Classification	6.42 (1.00)	5.12 (0.79)	5.03 (0.82)	5.35 (1.16)	5.02 (1.22)	5.18 (1.15)	4.87 (1.05)	-	-
3b									
CID-R	2.35 (0.63)	2.19 (0.51)	2.17 (0.60)	1.53 (0.46)	1.58 (0.53)	1.07 (0.30)	1.21 (0.37)	-	-
4									
Classification	7.45 (1.49)	-	-	6.17 (1.57)	-	-	-	5.48 (1.34)	-

Note. Standard errors are in parentheses. A dash indicates that this condition was not included in the experiment.

Table 6.

Mean effect of study duration (long condition minus brief condition) on priming (ms) and recognition (hit rate) for each new item RT quartile (SE in parentheses) in Experiments 1-3.

		Quartile			
		1	2	3	4
Classification task					
	Participants in each quartile	23	23	23	24
	M new item RT	590	871	1047	1379
		(17)	(14)	(9)	(43)
	M Priming (long – brief)	0.69	14.05	6.41	50.43
		(8.41)	(11.98)	(18.18)	(18.67)
	M Recognition (long – brief)	.12	.10	.09	.10
		(.02)	(.01)	(.02)	(.01)
CID-R task					
	Participants in each quartile	18	18	18	17
	M RT new items	2032	2475	2963	3735
		(48)	(27)	(31)	(111)
	M Priming (long – brief)	40.79	93.42	72.16	137.59
		(26.71)	(52.05)	(67.00)	(77.59)
	M Recognition (long – brief)	.08	.07	.04	.08
		(.02)	(.03)	(.03)	(.04)

Note. The brief condition comprises the data combined across the 250 ms condition from Experiments 1 to 3. The long condition comprises the data combined across the 2250 ms conditions in Experiments 1 and 2, and 1900 ms (longest) condition in Experiment 3.

Table 7.

Mean recognition endorsement rates for high confidence (HC) and low confidence (LC) responses in Experiment 4 (SE in parentheses).

Stimulus type	Response type			
	Old HC	Old LC	New LC	New HC
Brief old	0.45 (0.02)	0.19 (0.01)	0.18 (0.02)	0.18 (0.02)
Long old	0.54 (0.02)	0.17 (0.01)	0.14 (0.01)	0.14 (0.02)
New	0.14 (0.02)	0.14 (0.02)	0.28 (0.02)	0.44 (0.03)

Table 8.

Mean recognition endorsement rates for high confidence (LC) and low confidence (LC) responses computed as a function of response time (RT) on the priming measure (in ms) in Experiment 4 (SE in parentheses).

Stimulus type	Classification RT	Recognition response type			
		Old HC	Old LC	New LC	New HC
Brief old fastest	584 (23)	0.41 (0.02)	0.21 (0.02)	0.19 (0.02)	0.18 (0.02)
Brief old slowest	938 (45)	0.48 (0.02)	0.16 (0.01)	0.17 (0.02)	0.19 (0.02)
Long old fastest	591 (23)	0.52 (0.03)	0.18 (0.02)	0.16 (0.01)	0.14 (0.02)
Long old slowest	943 (44)	0.57 (0.03)	0.16 (0.02)	0.12 (0.01)	0.15 (0.02)
New fastest	592 (24)	0.14 (0.03)	0.14 (0.01)	0.29 (0.03)	0.43 (0.03)
New slowest	943 (45)	0.14 (0.02)	0.14 (0.01)	0.28 (0.02)	0.45 (0.03)

Table 9.

Goodness-of-Fit Values for the Single-System Model

Experiment	*N*	*p*	*q*	ln(*L*)	AIC	BIC
1a	31	6	5862	-42515.79	85403.58	86645.36
1b	24	6	4470	-38753.24	77794.47	78716.81
2a	31	6	5851	-43324.37	87020.74	88262.17
2b	18	6	3207	-28155.77	56527.55	57183.44
3a	29	10	11956	-94095.95	188771.89	190914.70
3b	26	10	5139	-45891.169	92302.34	94003.94
4	40	6	10372	-77790.80	156061.60	157809.03

Note. p is the number of free parameters in the model. *N* is the number of participants who were modelled (who had nonzero hit, miss, false alarm and correct rejection responses). The AIC (Akaike, 1973) and BIC (Schwarz, 1978) provide a measure of fit to the data. The AIC is calculated as follows: AIC =-2ln(*L*) + 2*Np*, where *Np* is the total number of free parameters for each fit. The BIC is calculated as follows: =-2ln(*L*) + *Np*ln(*q*), where *q* is the total number of data points fit.

Table 10.

Mean (and SD) of the Estimated Parameters of the Single-System Model

	Experiment						
Parameter	1a	1b	2a	2b	3a	3b	4
$\mu_{40\ ms}$	-	-	-	-	0.21 (0.20)	0.23 (0.26)	-
$\mu_{150\ ms}$	-	-	-	-	0.88 (0.37)	0.88 (0.49)	-
$\mu_{250\ ms}$	1.82 (0.35)	1.50 (0.59)	1.73 (0.48)	1.35 (0.94)	1.07 (0.49)	1.01 (0.57)	1.03 (0.57)
$\mu_{350\ ms}$	-	-	-	-	1.19 (0.53)	0.96 (0.53)	-
$\mu_{500\ ms}$	-	-	-	-	1.24 (0.48)	1.09 (0.67)	-
$\mu_{1900\ ms}$	-	-	-	-	1.40 (0.57)	1.16 (0.71)	-
$\mu_{2000\ ms}$	-	-	-	-	-	-	1.28 (0.70)
$\mu_{2250\ ms}$	2.24 0.47)	1.87 (0.59)	2.19 (0.61)	1.57 (0.91)	-	-	-
C	1.03 (0.38)	0.90 (0.39)	1.09 (0.35)	0.85 (0.58)	0.67 (0.39)	0.75 (0.44)	0.66 (0.47)
b	892.18 (262.42)	2429.76 (454.06)	967.39 (342.62)	2403.87 (431.00)	1052.41 (296.68)	3230.71 (589.03)	769.23 (213.23)
s	18.73 (26.37)	197.61 (74.27)	26.51 (25.40)	231.22 (179.81)	47.10 (50.90)	278.38 (178.32)	9.60 (18.10)
σ_p	243.21 (96.32)	871.87 (150.27)	309.42 (195.74)	963.34 (228.49)	394.92 (193.26)	1014.10 (135.48)	224.34 (94.57)

Note. Fixed parameters of the model are not shown in the table, but were set at $\sigma_f = \sqrt{0.5}$, $\sigma_r = \sqrt{0.5}$, $\mu_{new} = 0$, $M(e_p) = 0$, $M(e_r) = 0$, across all experiments, as in previous applications of the model (e.g., Berry et al., 2012, 2014). A dash indicates that the parameter was not used in the experiment.

Table 11.

Mean proportion of old judgments as a function of RT on the priming measure (in ms) in Experiments 1-3 (SE in parentheses).

		Stimulus Type					
		Brief		Long		New	
		M RT	p("old")	*M* RT	p("old")	*M* RT	p("old")
Classification task	Fastest	714	.73	704	.82	741	.21
		(20)	(.01)	(19)	(.01)	(22)	(.01)
	Slowest	1165	.72	1138	.84	1212	.18
		(43)	(.01)	(39)	(.01)	(44)	(.01)
CID-R task	Fastest	1734	.69	1668	.74	1957	.27
		(61)	(.02)	(59)	(.03)	(68)	(.02)
	Slowest	3225	.61	3123	.69	3623	.20
		(93)	(.02)	(88)	(.03)	(97)	(.02)

Note. The fastest and slowest categories were defined by a median split. The brief condition comprises the data combined across the 250 ms condition from Experiments 1 to 3. The long condition comprises the data combined across the 2250 ms conditions in Experiments 1 and 2, and the 1900 ms (i.e., the longest) condition in Experiment 3.

Figure 1. Mean priming effect and recognition hit rate as a function of study duration in Experiment 1a (panels a and b), Experiment 2a (panels c and d), Experiment 3a (panels e and f), and Experiment 4 (panels g and h). A classification task was used to measure priming in these experiments. The bars denote the experimental data and error bars denote 95% confidence intervals. Open circles denote the mean of the expected model results across participants. ns = no significant difference. * $p < .05$. ** $p < .01$. *** $p < .001$.

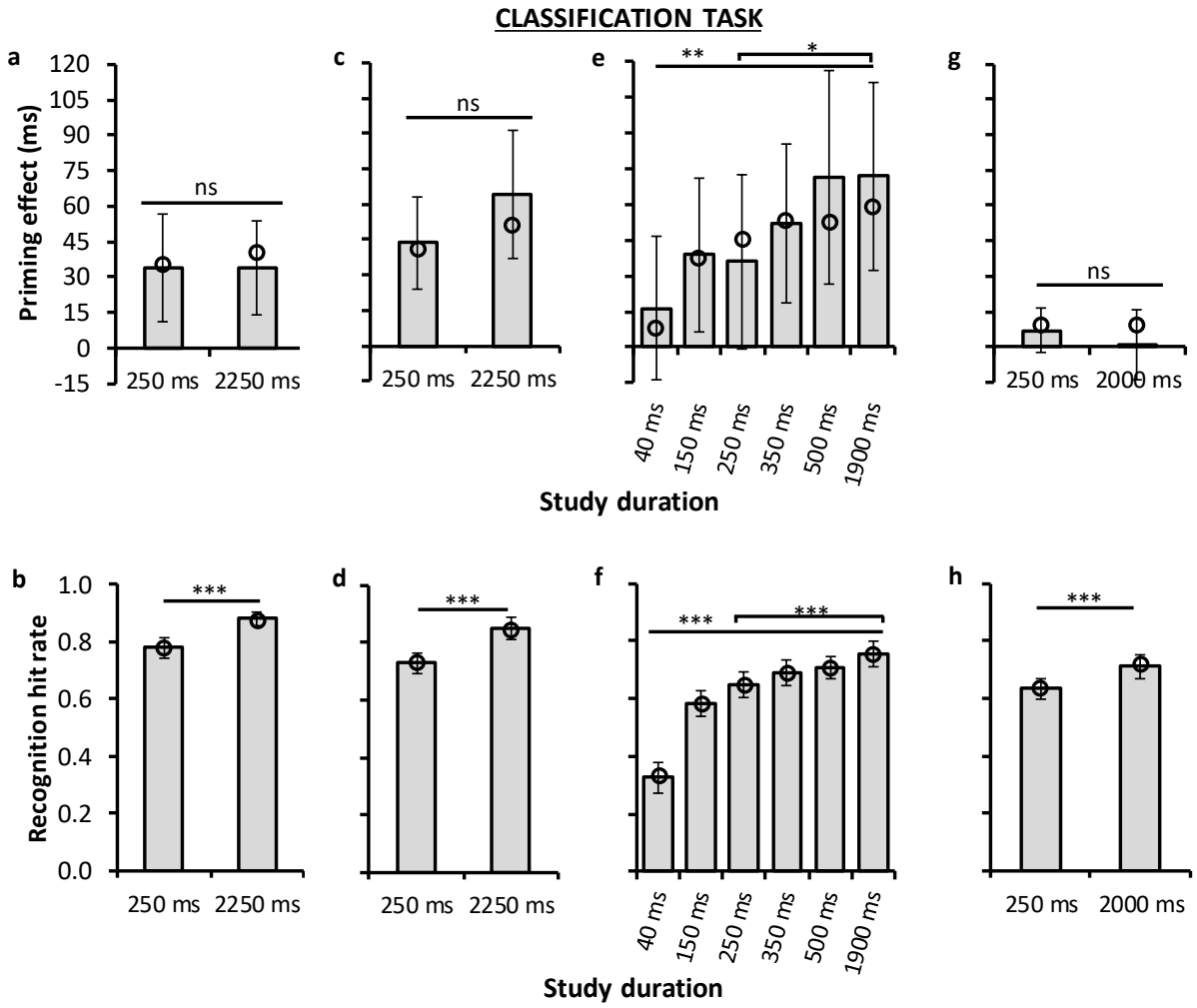

Figure 2. Results relevant to Predictions 2 and 3 of the single-system model. The top row shows the results from Experiments 1a (panel a), 2a (panel c) and 3a (panel e) that used the classification task to measure priming. The bottom row shows the results from Experiments 1b (panel b), 2b (panel d) and 3b (panel f) in which a CID-R task was used to measure priming. The bars denote the experimental data and error bars denote 95% confidence intervals. Open circles denote the mean of the expected model results across participants. Relevant to Prediction 2 is the differences between the mean RT to items in the priming task for misses and hits (M – H) and the difference in the mean RT for correct rejections and false alarms (CR – FA). Relevant to Prediction 3 is the difference in mean priming effect and the priming effect for items judged new (P – PJn) (where the priming effect for items judged new is calculated as the mean RT for correct rejections minus the mean RT for misses). The old item stimuli condition is given in parentheses (e.g., 250 ms).

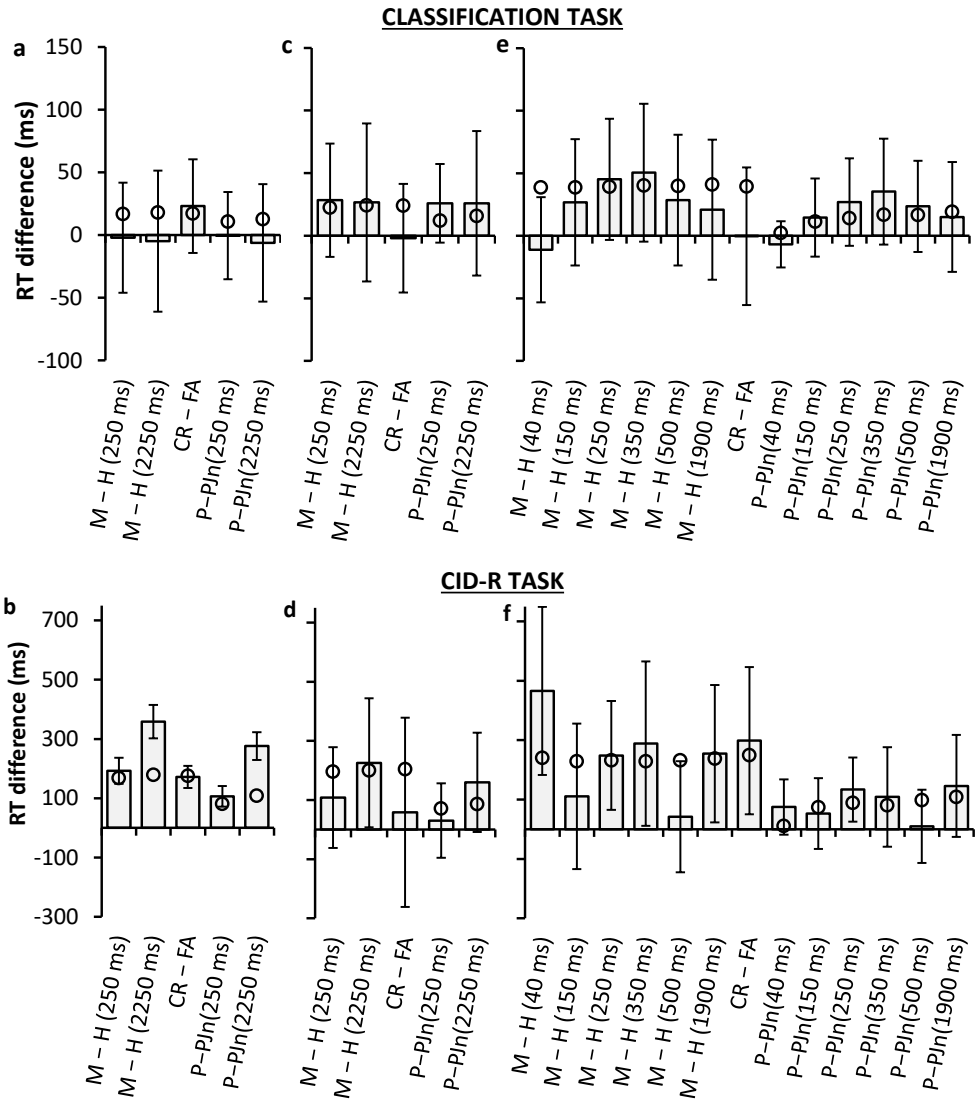

Figure 3. Mean priming effect and recognition hit rate as a function of study duration in Experiment 1b (panels a and b), Experiment 2b (panels c and d), and Experiment 3b (panels e and f). A CID-R task was used to measure priming in these experiments. The bars denote the experimental data and error bars denote 95% confidence intervals. Open circles denote the mean of the expected model results across participants. * $p < .05$. ** $p < .01$. *** $p < .001$.

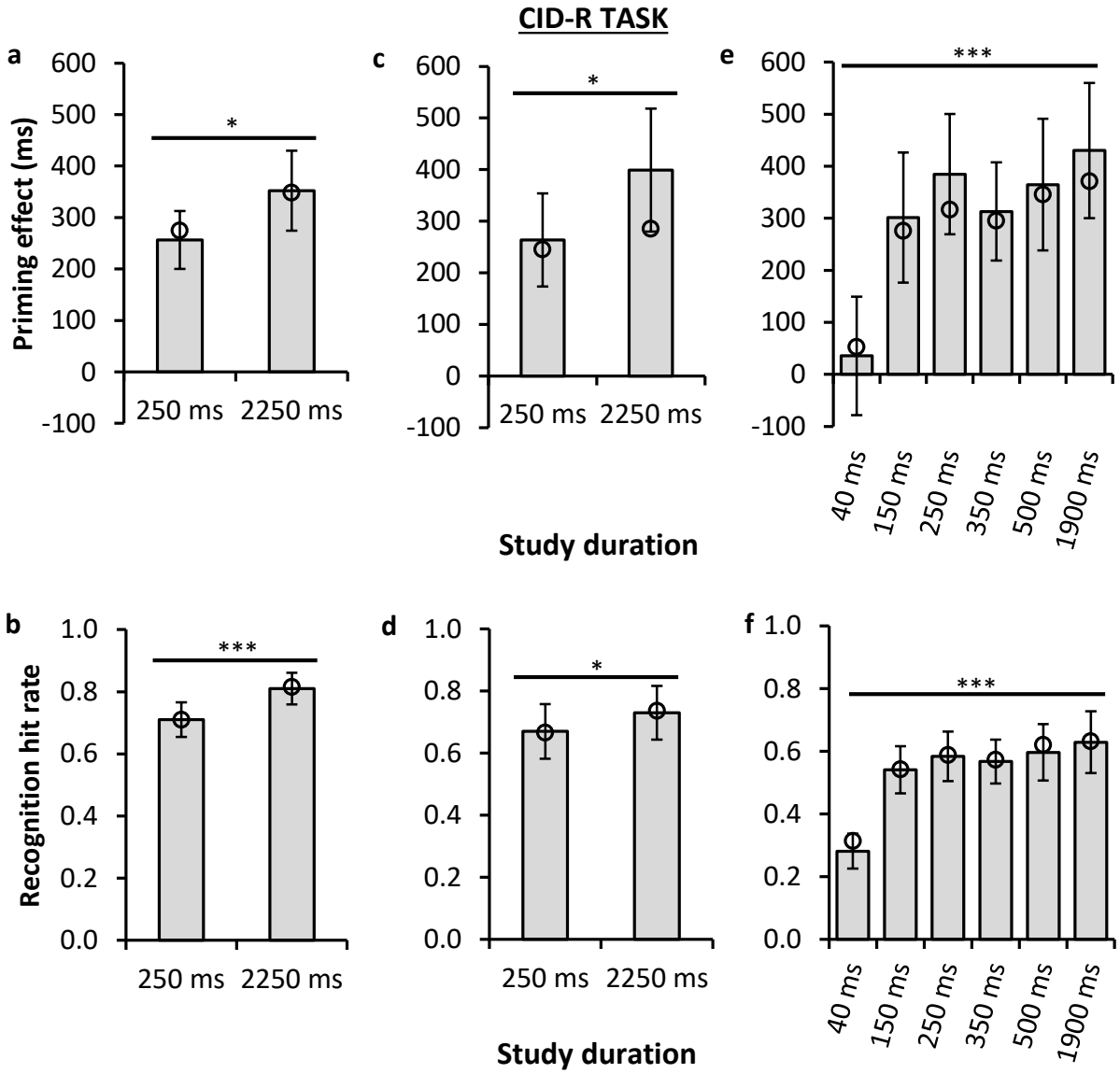

Figure 4. Results relevant to Predictions 2 and 3 of the single-system model (Experiment 4). The bars denote the experimental data and error bars denote 95% confidence intervals. Open circles denote the mean of the expected model results across participants. M – H = mean RT for misses minus mean RT for hits; CR – FA = mean RT for correct rejections minus mean RT for false alarms; P – PJn = Mean priming effect minus the priming effect for items judged new (where the latter effect is calculated as mean RT for correct rejections minus the mean RT for misses). The old item stimuli condition is given in parentheses.

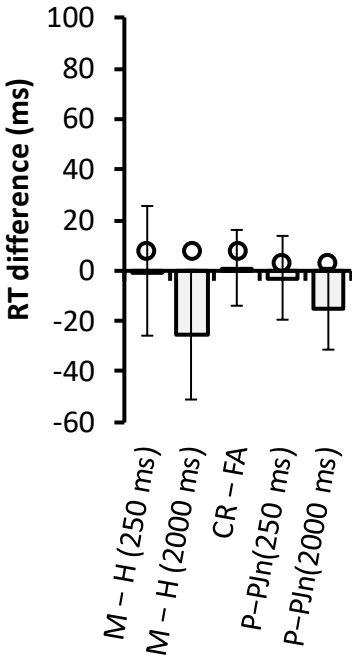

Figure 5. Results from a Bayes factor replication test (Verhagen & Wagenmakers, 2014) applied to the classification task data from Experiment 4. The dotted line represents the prior distribution, based on the effect size in Voss and Gonsalves (2010) (priming for brief minus long). The solid line shows the posterior after taking into account the data from Experiment 4. The white circle and gray square indicate the ordinate of the prior and posterior distribution, respectively, at the null hypothesis that the effect size is zero.