Why many priming results don't (and won't) replicate: A quantitative analysis

Michael Ramscar, Cyrus Shaoul & R. Harald Baayen
Eberhard Karls University, Germany

Running Head: Why priming results don't replicate

Corresponding author:
Michael Ramscar
Department of Quantitative Linguistics
Eberhard Karls Universitt Tbingen
Germany
Email: michael.ramscar@uni-tuebingen.de

Abstract

In what is now a controversial study, Bargh, Chen & Burrows (1996) reported that priming undergraduate students with words related to old age subsequently slowed the rate at which the students walked down a hall. Recent failures to replicate this and similar findings in the social-priming literature has prompted a heated debate, and led to widespread efforts to replicate a wide range of priming results. By means of a detailed quantitative analysis of Bargh et al's methods and materials, we show that both direct and conceptual efforts to replicate many specific priming effects can be expected to fail over time, between languages or across different age groups, regardless of the psychological veracity of their original results, simply because the learned behaviors that these studies seek to prime does not exhibit the invariance that typically supports direct replication in other areas of science. The dynamics of learning – and related cultural and linguistic change – present serious challenges to the scientific study of priming, yet these factors are overwhelmingly ignored by researchers. We describe a range of techniques that enable these factors to be objectively measured, and their influence more formally estimated.

Why many priming results don't (and won't) replicate

In experimental psychology, the word 'priming' is often used to describe a range of findings in which subjects' behavioral responses to a given task are manipulated by means of exposure to other meaningful stimuli such as words and pictures. For example, a subject might be asked to read a prime word and then judge whether a letter string is either a word or nonword. If the prime is linguistically related to the target in a language spoken by the subject (e.g., when the prime is 'doctor' and the target is 'nurse'), the target is likely to be more quickly classified as being a word (Meyer & Schvaneveldt, 1971). This result is assumed to occur because as subjects make lexical decisions, the prime word serves as an informative cue to the target string. That is, some aspect of the stimulus properties of the prime word reduces the subject's uncertainty about the correct response on a given trial, and this makes the task easier. Further support for the idea that priming influences the uncertainty associated with subsequent behavior – such as making a lexical decision – comes from the finding that although the prime 'doctor' leads to quicker classifications of the target 'nurse,' this classification will be made even faster if the prime word is 'nurse' itself (see e.g., Forster & Davis, 1984).

While the informativity of primes about targets is straightforward and even quantifiable in studies examining straightforward lexical relationships, such as lexical decision tasks (Milin, Ramscar, Cho, Baayen & Feldman, 2015), or lexical paired associate learning (PAL; Ramscar, Hendrix, Shaol, Milin & Baayen, 2014; Ramscar, Hendrix, Love & Baayen, 2013), in other research paradigms, the relationships between the materials that are used to prime subsequent behaviors and those behaviors themselves are more abstract. For example, in a famous study of social priming, Bargh, Chen & Burrows (1996) presented evidence indicating that prime words associated with 'the elderly stereotype' subsequently slowed the rate at which subjects walked along a corridor. In this study (Bargh et al, 1996, Experiment 2), subjects completed a scrambled sentence task (Srull & Wyer, 1979) in which they rearranged sentences in order to make them grammatical. The sentences contained either words relating to the elderly stereotype or neutral words. After completing the task, Bargh et al. surreptitiously measured the time it took their subjects to walk along a corridor after exiting a laboratory room. Subjects primed with elderly words walked slower than subjects in the neutral condition.

Although the kind of general effects that result from straightforward lexical priming paradigms have proven to be relatively robust (that is, these effects replicate on average over large sets of items), the reliability of more abstract (and specific) priming paradigms such as that employed by Bargh et al have recently been the subject of much debate, with a number of research teams (e.g. Doyen, Klein, Pichon & Cleeremans, 2012; Pashler, Harris, & Coburn, 2011) reporting that Bargh et al's results failed to replicate on other groups of subjects given the same priming procedures.

Doyen et al's failure to replicate these results, which coincided with a series of high-profile fraud cases among priming researchers (Yong, 2012) led to considerable controversy: In a much discussed article reflecting on these matters, Kahneman (2012) described priming studies as, "the poster child for doubts about the integrity of psychological research". Acknowledging that "priming effects are subtle and that their design requires high-level skills", he advised researchers to "organize an effort to examine the replicability of priming results, following a protocol that avoids the questions that have been raised and guarantees credibility among colleagues outside the field."

These events also resulted in the appearance of special issues of several journals that were dedicated to the perceived "crisis" in the replicability of psychological results. Pashler & Wagenmakers (2012), the editors of one such issue, echoed Kahneman's call for more replication, warning that psychologists have "found ourselves in the very unwelcome position of being. . . the public face for the replicability problems of science in the early 21st century." Similar commentaries (e.g., Asendorpf, et al., 2013) also argued for a drastic increase in the number of replication studies – and in particular, direct replication studies, which seek to repeat experiments without any changes to the materials or methods (Cesario, 2014; Simons, 2014) – claiming that such an effort would be required in order to rehabilitate the credibility of psychological science. These and other similar calls have led in turn to the publication of yet more high profile failures of direct replications (e.g., Harris, Coburn, Rohrer & Pashler, 2013; Shanks et al, 2013).

In what follows, we show that as a consequence of the very methods that many priming studies employ, increasing the number of direct replications that are conducted by researchers will lead to a waste of time and resources rather than an increase in scientific understanding, simply because a range of social, linguistic and contextual factors makes the direct replication of many studies impossible. As we will show, because of these readily quantifiable factors, finding that a result does not replicate is highly unlikely, in itself, to bring any scientific clarity to whatever theoretical issues that prompted the initial study. The reasons for this can be simply stated: The informativity of the prime-target relationships embodied in the materials employed in priming studies is something that subjects have to learn, and as such, the nature of these relationships is a product of the environments and cultures to which subjects are exposed. Accordingly:

1. Because relationships between primes and targets are subject to change as culture changes, even within a given culture, "valid" priming results ought not to be expected to replicate across time

2. Because the structure of cultural technologies like languages varies across cultures, priming effects cannot be expected to replicate across them, and so priming results ought not to be expect to survive translation into another tongue even in cases where cultures are closely related

3. Because priming effects are the product of learned relationships, and because learning is a lifelong, discriminative process, the informativity of primes constantly – and systematically – changes with experience, and so priming results ought not to be expected to be invariant over the lifespan, even if culture is held constant

To illustrate these points, we will focus on the well-known Bargh et al (1996), Experiment 2 (in which participants completed a scrambled sentence task and then were measured for walking speed) because its fame, its typicality and the fact that numerous replications of it (and failures to replicate) have made an ideal exemplar of the broader category of priming studies at the heart of the present controversy. We note, however, that given the widespread degree to which social-priming research relies for its effects on linguistic and symbolic primes, the problems we highlight will apply to most, if not all, social priming effects (and to some degree, to all other priming effects in psychology as well).

To establish these facts, and to offer some quantitative means that can help ameliorate some of their negative implications, we describe a range of methods that offer a means for analyzing the likelihood that an effect might replicate in other contexts given a set of linguistic or symbolic primes. As well as improving our understanding of the replicability of priming effects, these methods can offer insight into the theoretical and empirical foundations of priming, and the mechanisms underlying priming effects.

1. Why priming studies cannot be expected to replicate over time

Although the exact nature of the relationship between social priming and other forms of psycholinguistic priming has yet to be fully determined (see Neely, 1991 for a review), in the case of Bargh et al (1996) Experiment 2, the relationship between psycholinguistic priming and social priming can be assumed to be close, simply because the set of primes employed and manipulated in the study were words. Given this, it follows that the effect reported must be some function of the lexical properties of the prime set.

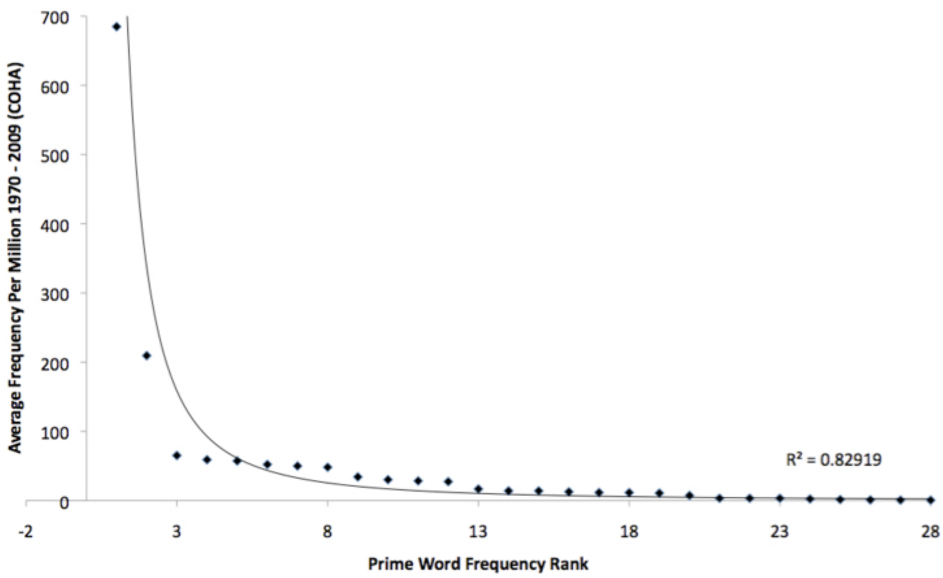

Figure 1: Items used in Bargh et al. (1996) Experiment 2 plotted by frequency and rank based on their 2000-2010 distribution in COHA (the trend line plots the fit to a power function; see Zipf, 1949).

To introduce these lexical properties, Figure 1 plots the prime words' frequencies in the 2000-2010 sample of the Corpus of Historical American English (COHA; Davies, 2012). As can be seen, even this small set of items exhibits the skewed distribution characteristic of linguistic samples (Baayen, 2001; Zipf, 1949). As a result, one prime item is actually as frequent on average as the rest of the items combined.

It is worth highlighting another property of linguistic distributions in relation to this: The distributions of words and word n-grams in natural languages are not only Zipf distributed at every level of analysis (Baayen, 2001), but these distributions also exhibit a statistical property known as burstiness. In language use, frequency is not a fixed property of a word, but rather, whereas the frequencies of a relatively few common words will tend to reflect their average frequency in most texts, the frequencies of most words are over-represented (based on their average across texts) in a small number of texts and under-represented in all others (Church & Gale, 1995; Katz, 1996). Thus, whereas the average frequency of the word and will be fairly consistent across 100 randomly sampled texts, the word n-gram is unlikely to be found in any of them. And in texts where the n-gram does occur, it will do so far more often than one might expect given its average frequency across a large sample of texts.

Similarly, if we were to survey 100 randomly sampled individuals, their shared experience of the word will be fairly consistent, whereas few of them may have encountered a specific word n-gram, if any have at all. It thus follows that the more frequent a lexical prime item is, the more reliable its priming effects can be expected, on average, to be (and the less frequent a prime item is, the less reliable its priming effects can be expected to be).

These considerations indicate in turn that, since any given set of linguistic primes will be Zipf distributed (see Figure 1), it follows most, if not all of the priming that is afforded by any given set of primes is likely to be the result of only a small subsection – perhaps even just one – of the members of any given prime set. One immediate point that these considerations underline is the importance of care when reporting priming effects. Bargh & Chartrand (1999, p466) describe the original Bargh et al experiment as follows:

"In a first task, participants were primed (in the course of an ostensible language test) either with words related to the stereotype of the elderly (e.g., **Florida, sentimental, wrinkle***) or with words unrelated to the stereotype. As predicted, participants primed with the elderly-related material subsequently behaved in line with the stereotype–specifically, they walked more slowly down the hallway after leaving the experiment."*

Yet when summed, the frequency of Florida, sentimental, wrinkle in the two decades prior to Bargh et al's experiment (i.e., over the lifespan of the participants tested) amounted to just 45 instances per million words of text. By contrast, as Figure 1 shows, the frequency of one item in the Bargh et al primes amounts to around half of the summed average rate of occurrence of all of the words in the words in the prime set: That word, which occurs on average 750 times in each million words of English, is *old*.

The word *old* is, of course, used frequently in English as a descriptor for elderly people: in every decade since 1810, the noun most likely to follow *old* in American English is *man*, with *woman* next (*lady* has also consistently been a likely occurrence given *old*;

Davies, 2009). It thus follows that once the burstiness of word frequencies is also taken into account, the contribution of the word *old* to the average effect Bargh et al's observed across their subjects is likely to have been far far greater than that of the words *Florida, sentimental,* and *wrinkle.* Indeed, given the relatively small size of the effect reported, and the relative frequencies of the words, the influence of the primes *Florida, sentimental,* and *wrinkle* on the average walking speed of undergraduates is likely to have been minimal. It thus follows that a more accurate description of the manipulation in the Bargh et al experiment would be:

"participants were primed either with the word **old** *(which is closely related to the stereotype of the elderly) or with unrelated words."*

The lexico-statistical facts that we have described also suggest a possible explanation for the priming effect observed by Bargh et al., as well as a means for quantifying its influence. In much the same way that the priming relationship between 'doctor' and 'nurse' is a function of the distributional properties of English, Ramscar, Matlock & Dye (2010a) showed that priming effects attributed to other loosely-defined conceptual structures – 'embodied motion schemas' (Boroditsky & Ramscar, 2002; Matlock, Ramscar & Boroditsky, 2005) – can be explained in more quantitative terms by analyzing the way words can be expected to prime other words based on their distribution in large samples of language.

For example, Matlock et al. (2005) report that after judging whether sentences such as *"Eight pine trees run along the edge of the driveway"*, were intelligible in English, subjects subsequently showed a future bias when answering the ambiguous temporal question, *"Next Wednesday's meeting has been moved forward two days. What day is the meeting now that it has been rescheduled?"*. If sentences featured eight or twenty trees, subjects were more likely to respond, "Friday," whereas eighty trees induced no bias.

Subsequent empirical and statistical analyses (Ramscar et al 2010a) revealed that this pattern of priming – and the priming induced by a further set of primes – correlated well with the way that time-words (e.g., years, days) are represented in the distributions following the words used as number primes in the Corpus of Historical American English (Davies, 2009; a 450 millon-word sample of American English text and speech). The direction of the priming results was also explained by the distribution of English words: the manipulation induced a future bias in subjects answering 'the ambiguous meeting question', and analysis revealed that, in English, collocations of integer- and time-words occurs far more frequently discourse related to the future than to the past.

To return to the Bargh et al study: Given that old is a frequent collocate of words like man and woman, and given that collocation has been shown to be a reliable predictor of priming (indeed, it is to our knowledge the only objective method for quantifying these priming effects), these findings suggest that analyzing the collocational properties of Bargh et al's lexical primes provides a means for estimating the degree to which they 'old' can be expected to prime ideas about elderly people (i.e., elderly "trait concepts and stereotypes") across contexts, and to examine how the nature of this priming might behave as contexts change. To this end, Figure 2 plots the summed frequency of old man and old woman in COHA in the period 1970-2009. As can be seen, the frequency with which these noun phrases were used in American English decreased by around 25% over this period.

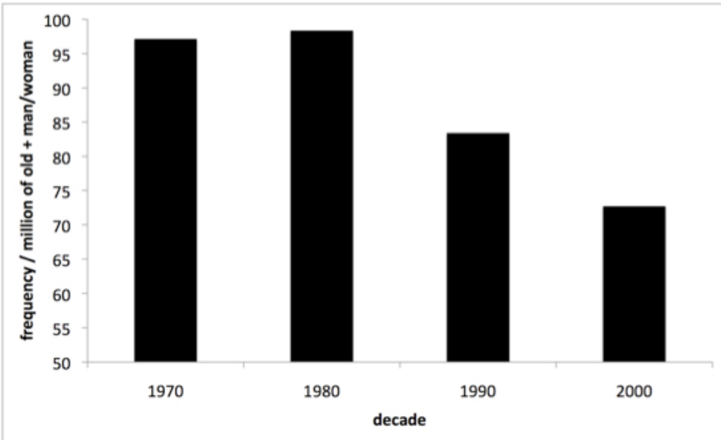

Figure 2: Summed frequencies of the phrases old man and old woman in the period 1970-2010 (in COHA, Davies, 2009). Man and woman were the nouns most likely to follow old in American English in this period.

Figure 3 then plots the distribution of a larger set of terms that relate to people (the intended targets of the priming, i.e., man, woman, folks, etc) following the word old in COHA in the same period. As can be seen, although the number of elderly people in the US population increased dramatically in the period 1970-2009 (Ramscar et al, 2014, see below), the likelihood of old being used as a descriptor for people (e.g., old man, old woman, old folks) actually decreased by more than 25% in this period (paired log frequency difference between the 1970s and 2000s for the items in Figure 1 panel C: $t(9)=1.833$, $p=0.002$). This suggests that the degree to which old might be expected to prime concepts about elderly people is likely to have decreased as well.

An analysis of the full set of prime words serves to illustrate of the behavior of linguistic distributions over time. Table 1 presents a comparison of the COHA frequencies of all of the items from the 1970s (sample = 23.8 million words) with the 2000s (sample = 29.5 million words), and reveals that while the rank position and frequency of old remained stable from 1970 to 2010, other items in the prime set experienced greater degrees of distributional volatility in this time-frame: Withdraw was three times more frequent during the 1970s than the 2000s, whereas Florida was around twice as frequent in the 2000s than the 1970s (it is possible that the former relates to the Vietnam war, which was prominent during the earlier period, and the latter to the 2000 Presidential election, which was prominent during the later period). From the perspective of the influential Distributional Hypothesis (Firth, 1957; Landauer & Dumais, 1997; McDonald & Ramscar, 2001; Jones & Mewhort, 2007; Johns & Jones, 2010; Ramscar et al, 2010; Ramscar, Dye & Klein, 2013), which proposes that many aspects of word meaning are determined by the linguistic contexts in which words occur, the meanings of withdraw and Florida

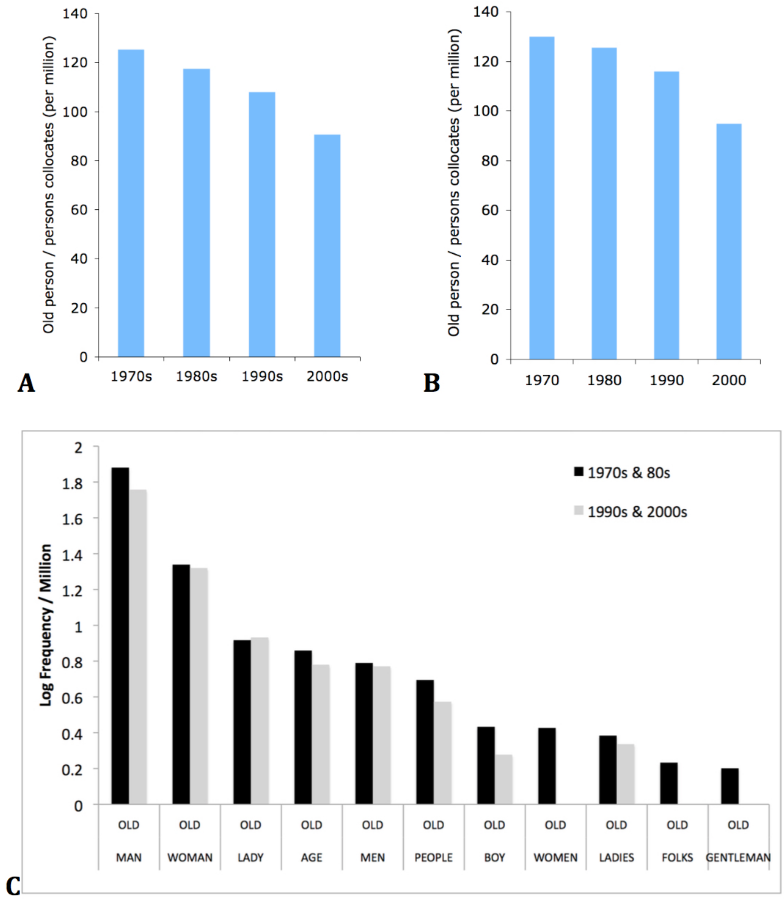

Figure 3: Cumulative collocations frequencies (per million) for people terms (woman, folks, etc) following old, 1970-2010. **Panel A** shows the cumulative frequency of the 10 most frequent collocates in a 3 word window following 'old,' and **Panel B** the cumulative frequency of the 30 most frequent words immediately following 'old.' (A minimum information criterion was applied to the search, to exclude function words such as as, and, etc).**Panel C** plots the log frequencies of a larger set of old…people pairs in the 1970 and 80ss as compared to the 1990s and 2000s.

would be considered to have changed in this period. And given that language learning continues throughout the lifespan (Ramscar et al, 2013c, 2014; Keuleers et al, 2014), it follows that it is likely that subjects sampling the distribution of American English in these different timeframes will have learned to form different expectations in the face of these words, and that these different expectations might influence the effects of their being primed with these words in behavioral tasks.

Assessing the impact that these and any other distributional changes may have had on the priming potential of the full set of Bargh et al. items is, however, less straightforward than for old. Most of the other primes – e.g, ancient, grey – do not occur as collocates of words relating to adults in the way that old does (indeed, in the distribution

	Frequency / Million 1970s	Frequency / Million 2000s	Frequency relationship (1970s/2000s)
OLD	756.31	636.04	1.19
ALONE	220.79	196.48	1.12
WORRIED	53.35	63.06	0.85
ANCIENT	53.05	63.77	0.83
TRADITIONAL	51.37	65.61	0.78
CAREFUL	51.33	51.39	1.00
CONSERVATIVE	49.18	35.11	1.40
FLORIDA	37.36	69.54	0.54
LONELY	34.25	26.49	1.29
RETIRED	31.60	39.52	0.80
BITTER	31.17	25.95	1.20
WISE	28.36	26.90	1.05
DEPENDENT	19.82	12.45	1.59
WITHDRAW	18.22	6.21	2.93
RIGID	17.00	10.62	1.60
HELPLESS	16.28	11.13	1.46
GREY	15.27	11.19	1.36
CAUTIOUS	13.88	9.94	1.40
STUBBORN	10.85	10.79	1.01
SENTIMENTAL	8.29	6.24	1.33
COURTEOUS	4.50	2.07	2.18
OBEDIENT	3.37	3.12	1.08
BINGO	2.27	4.04	0.56
GULLIBLE	1.64	1.22	1.34
WRINKLE	1.47	2.85	0.52
FORGETFUL	1.30	0.75	1.73*
KNITS	0.63	0.85	0.74*
SELFISHLY	0.55	0.54	1.02*

Table 1: The prime words employed by Bargh et al. (1996), sorted by their rank frequency per million in the 1970s, along with the corresponding frequencies in the 2000s, and the ratio of change in this time.

of English, one of the Bargh et al. primes – helpless – is far more likely to cue child and children). However, since it might be argued that even where items do not occur as direct collocates of targets, the full set of prime words might exert a collective priming influence on behavior as a result of higher-order distributional relationships to the target (see Jones, Kintsch & Mewhort, 2006 and Hutchison, Balota, Cortese & Watson, 2008 for opposing views on this point), we took the most frequent 5 words in the distribution following each of the Bargh et al primes in COHA in order to estimate the potential effect of these relationships (see Ramscar et al, 2010a).

To estimate the degree to which the higher-order distributional properties of the Bargh et al. items could be expected to prime elderly concepts over time, we then analyzed the relationship between the sets of words that are the higher-order relatives of the prime words and the word elderly using Latent Semantic Analysis (LSA), a corpus-derived measure of contextual similarity that has been shown to correlate well with empirical priming results, as well as human semantic similarity judgments, etc. (Landauer & Dumais, 1997). [1]

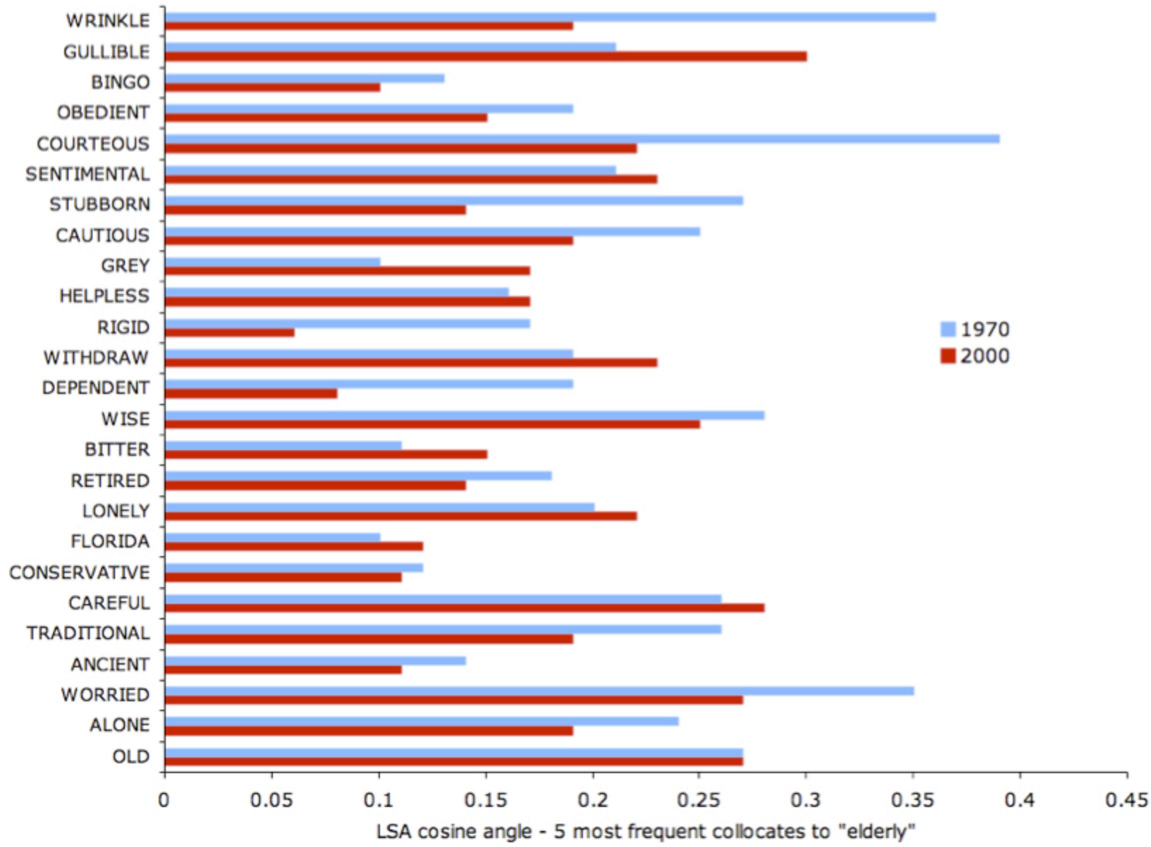

Figure 4: LSA cosine angles between the 5 most frequent direct collocates of each of the Bargh et al.(1996) items and the word 'elderly.' Greater cosine angles denote greater contextual similarities.

Figure 4 plots the LSA similarity scores for the set of higher-order collocates of each of the Bargh items and elderly. [2] As can be seen, if we take LSA similarity as an indicator of how strongly words prime other words through higher order relationships (i.e.,

[1]LSA similarity calculations were made using the LSA website at http://lsa.colorado.edu/. Forgetful, knits, and selfishly were excluded from these analyses because their low frequencies cause inferences drawn about their distributional properties to be highly unreliable, even when based on relatively large samples (Baayen, 2001). It is worth noting that, as we explained above, these same low frequencies mean that they will also be highly unlikely to promote reliable or predictable priming effects across a randomly sampled population of participants.

[2]Each word set comprised the 5 most frequent collocates in a 3-word window following each prime. A minimum mutual information score of 3 bits was used to eliminate uninformative but highly frequent

as an objective measure of at least some of the conceptual properties of those words), then the degree to which the Bargh et al prime words could be expected to prime elderly declined significantly in the period 1970-2010 (paired t(25)= -2.315, p=0.029, two-tailed).

A frequency-weighted re-analysis [3] confirmed that this was not an artifact of the properties of the less frequent primes, which as we described above, can be expected to influence the replicability of Bargh et al's findings less than higher frequency primes because of their lower frequencies (paired t(25)= -3.187, p=0.004, two-tailed).

To examine whether the predicted decline in the expected efficacy of the Bargh et al prime items in our first analysis replicated using a different analytic approach, a second LSA analysis was conducted in which the sets of Bargh et al prime collocates were pooled to form a single set, and then this set was compared to a wider range of 'elderly' target words (*senior, retired, elder, aged, old* and *elderly*).

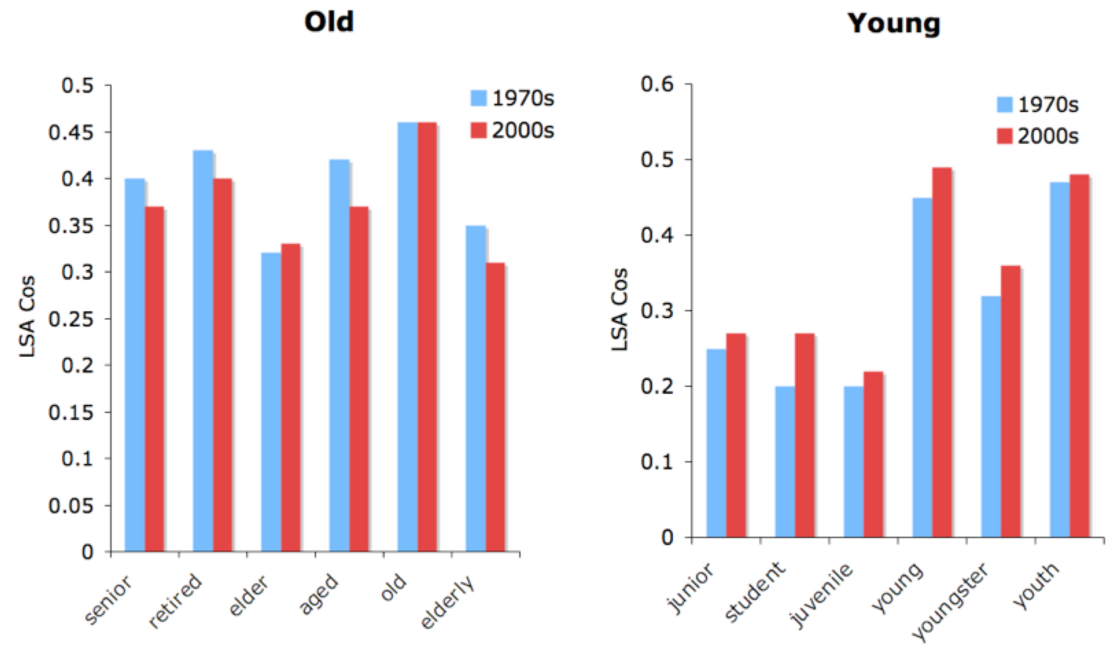

Figure 5: Average LSA cosine angles between the 5 most frequent direct collocates of each of the Bargh et al.(1996) items and 6 youthful and 6 elderly near-synonyms in the 1970s and the 2000s.

As the left panel of Figure 5 shows, predicted priming for four of the six words declined between 1970 and 2010 (paired t(6)=2.015, p=0.058, two tailed). Further, when we computed the average similarity between the same words primed by the Bargh et al.

content words, punctuation, etc. Counts were taken from the complete COHA sample for each relevant period.

[3] *Weighted Similarity = lsa similarity(word) × log(frequency word).*

prime set and a set of 'youthful' target words (*junior, student, juvenile, young, youngster*, and *youth*), we found the opposite pattern: predicted priming for youthful words increased between 1970 and 2010 (paired t(6)=-3.78, p=0.013, two tailed; Figure 5, right panel).

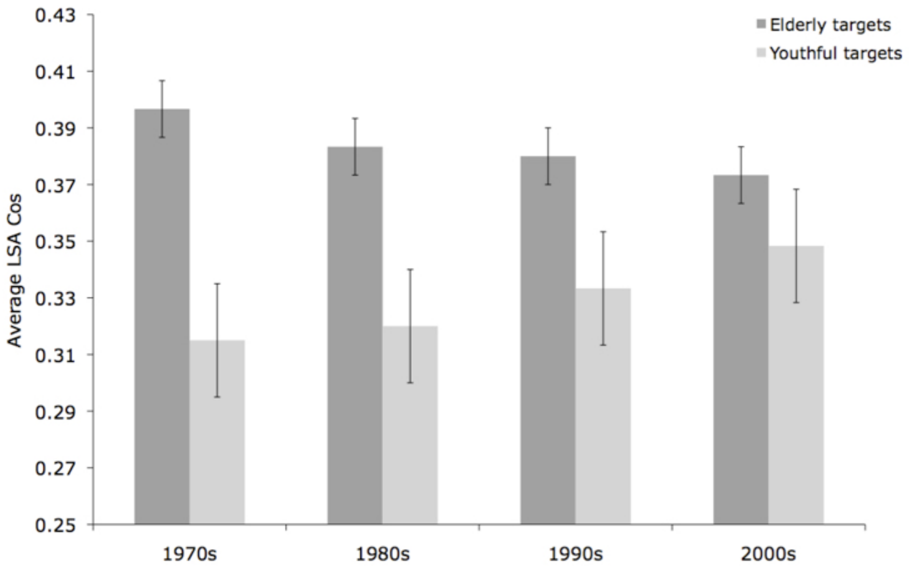

Figure 6: Average LSA cosine angles between the 5 most frequent direct collocates of each of the Bargh et al.(1996) items the youthful and elderly near-synonym sets, 1970 – 2010.

In summary, these analyses show that the degree to which old, by far the most frequent of the Bargh et al. primes, can be expected to prime elderly concepts has declined over the past forty years, and that the predicted priming for the full set of Bargh items has declined along with it. At the same time, the degree to which the overall set of items appears likely to prime youthful concepts has increased to the extent that, by this measure, the difference between the two is now no longer significant (Figure 6).

All of the analysis we have reported have provided objective reasons why we ought not to expect the Bargh et al (1991) Experiment 2 be to replicate in today's world. Moreover, these analyses offer fertile means for establishing why, in 2015, setting out to conduct a 'direct replication' of this experiment is something of a fools' errand. However, a question still remains: Might it possible to use the analytic methods we have described to construct a set of stimuli that would make a conceptual replication of the Bargh et al experiment a viable (and indeed, quantifiable) possibility?

While it seems clear that the change in usage of phrases such as old man and old woman has diminished the priming capabilities of the original Bargh et al stimulus set, perhaps this is because new words are now being used to elicit 'the elderly construct,' such that a new set of words with similar priming capacities, such as retiree or senior

	1970s	1980s	1990s	2000s
RETIRED	31,4	30,3	35,5	39,4
RETIREE	1,1	0,5	1,1	2,1
SENIORS	3,2	3,1	8,6	12,9
SENIOR CITIZENS	2,6	2,3	2,9	3,7
SENIOR CITIZEN	1,0	1,0	0,7	0,9
ELDERLY	24,9	24,6	23,1	20,6
AGED	16,2	15,4	12,4	12,9
PENSIONER	0,6	0,2	0,3	0,2
PENSIONERS	0,9	0,8	0,8	0,1
OLDER LADIES	0,0	0,0	0,1	0,1
OLDER LADY	0,1	0,1	0,1	0,1
OLDER ADULTS	0,0	0,1	0,4	1,3
OLDER PEOPLE	4,0	3,7	3,0	2,1
OLDER PERSON	0,2	0,4	0,4	0,6
OLDER MAN	5,6	4,2	4,3	5,4
OLDER MEN	1,9	1,6	1,7	1,6
OLDER WOMEN	1,2	2,0	1,3	2,2
OLDER WOMAN	2,4	2,7	2,9	3,9
TOTAL	97,3	93,0	99,5	110,0
OLD MAN	82,6	69,3	57,6	56,7
OLD WOMAN	14,6	29,1	25,8	16,0
OLD MEN	5,9	6,4	6,7	5,1
OLD WOMEN	2,5	2,4	2,5	1,3
OLD PERSON	0,5	0,3	0,4	0,2
OLD PEOPLE	5,3	4,6	4,2	3,3
OLD FOLKS	1,5	1,5	1,7	0,9
OLD AGE	7,8	6,7	6,3	5,7
OLD LADY	9,9	6,6	10,1	7,0
OLD LADIES	2,8	2,0	2,5	1,9
TOTAL	133,5	128,8	118,0	97,9

Table 2: Changes in the average frequency of a set of ‘old’ modified words and their semantic equivalents, decade by decade from 1970 - 2009.

citizen could be identified. That is, perhaps language usage is, in the limit, akin to a zero sum game, such that when people stop using old man and old woman as descriptors, it is because they are now using retiree and senior citizen etc. instead.

To examine this possibility, we analyzed the relative frequency of a set of ‘old’ modified words such as *old man* and *old woman* and a far larger set of their semantic equivalents, such as *retiree* and *senior citizen*, in each decade from 1970 – 2009 (Table 2).

Figure 7, which summarizes these data, reveals how, while to some extent, a linguistic action and reaction does appear to have occurred – decreases in the usage of some

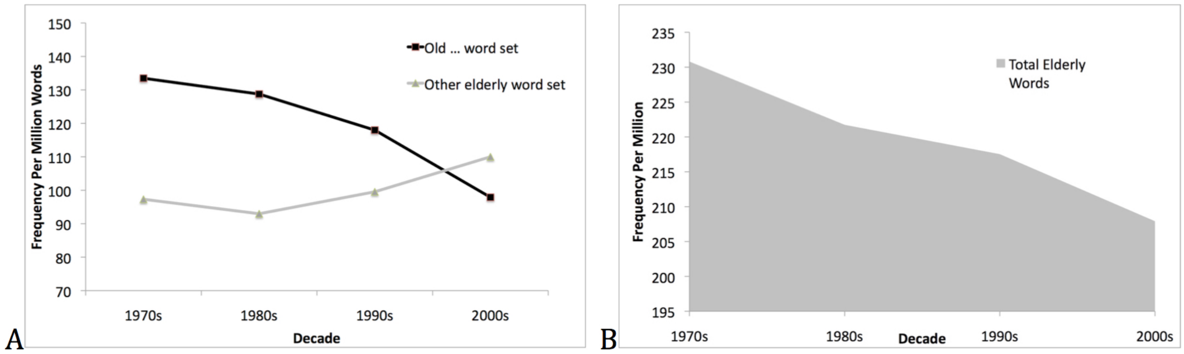

Figure 7: **Left panel**: Summed word frequencies for the set of 'old' modified words (*old man, old woman*) and a larger set of semantic equivalents (*retiree, senior citizen*) by decade from 1970 – 2009. **Right panel**: Summed frequencies for both sets of words in this period.

appears to be have been accompanied by increases in the use others – the overall trend is downwards. In the 1970s, the summed average frequency of all of our descriptors for the elderly was 231 instances per million words, whereas by the 2000s, this had dropped to 208 instances per million words. To put this change into some kind of context, the average frequency of the word elderly (the construct the Bargh et al primes were intended to elicit) across this period was around 23 instances per million words, which means that the change in the overall frequency of our set of 'elderly words' is equivalent to eradicating the word elderly itself from the English lexicon, while holding all other linguistic factors constant.

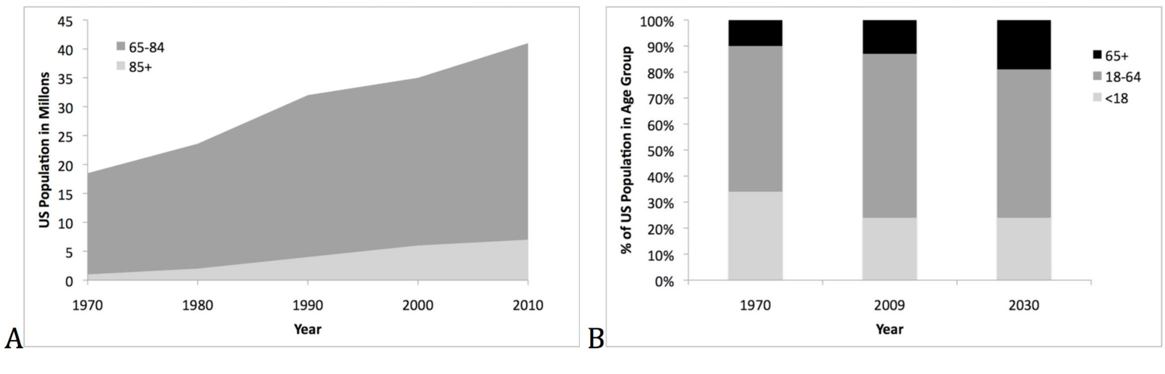

Figure 8: **Left panel**: Total population (in millions) of adults aged 65-84 and 85+ in the US 1970 – 2010. **Right panel**: The changing proportion of elderly adults in the US population (data in both plots from US Census Bureau).

It is worth reflecting on why this change might have occurred. Figure 8A plots the increase in the elderly population in the period 1970-2010, and Figure 8B shows how this relates to the total US demographic in relative terms. As can be seen, the number of elderly adults in the population has increased (and is increasing) in both absolute and relative terms. In information theoretic terms, it is something of a truism that as the proportion of the population that is elderly increases, describing any individual as elderly necessarily becomes less informative (Shannon, 1948). It thus appears that, to some degree, the change in the informativity of describing someone as elderly across this period has been reflected in the way that English is used.

A number of other things follow from these changes: First, the usage changes we have described can be expected to have impacted what these words actually prime, such that the fact that behavior that Bargh et al report twenty-year-old subjects exhibited in a priming study conducted in 1991 (Bargh, 2012c) do not materialize in direct replication attempts conducted some twenty years later (Pashler, Harris, & Coburn, 2011) tells us little about the validity of Bargh et al's original findings, simply because the lexical changes in the properties of the primes we have described offer ample reason to expect in advance that a valid priming effect from two decades ago will fail to replicate in today's twenty-year-olds. Moreover, when these lexical changes are taken together with the demographic changes plotted in Figure 8, it seems that the cultural changes they reflect are likely in turn to rule out any notion of a conceptual replication of Bargh et al's Experiment 2 as well, simply because it seems inevitable that the conceptual knowledge participants now bring to bear in experiments will also have changed.

Or, to put it another way, Bargh et al described the primes used in their study as being designed to evoke 'the elderly stereotype'; The data and analyses we have presented offer a great many reasons to believe that that this stereotype has changed significantly in the time that has elapsed since Bargh et al's study was originally conducted.

2. Why priming studies cannot be expected to replicate across languages

The publication of one particular failure to replicate Bargh et al's findings by Doyen, Klein, Pichon & Cleeremans (2012) generated a considerable amount of controversy among psychologists and other interested parties (e.g., Bargh, 2012, a,b,c; Yong, 2012; Kahneman,2012). Not only did Doyen et al. fail to find an effect of priming on walking speed in their attempted replication of the Bargh et al study, but they also reported that they could only succeed in obtaining a walking speed effect using Bargh et al's protocols if the experimenters actively expected the effect they were trying to detect (i.e., when Bargh et al.'s study was not conducted under double blind conditions). Doyen et al's findings indicated that Bargh et al's original results were 'clever Hans effects' (Yong, 2012). Doyen et al's article had been cited around 200 times up until June 2015. However, this much-discussed and analyzed 'failed replication' made at least one significant change to Bargh et al's methods that has thus far gone un-noticed in the burgeoning literature devoted to this debate: Doyen et al tested French-speaking subjects and their attempted replication of the Bargh et al study utililized French-language primes.

Doyen et al. appear to assume that the language a priming study is conducted in is not an important factor in research, or indeed, in replication (this assumption is widespread in the literature, see e.g., Gmez, Daz, & Marrero, 2011; Aveyard, 2014). However, although most influential social priming studies have been conducted in English, there is reason to believe that this assumption is mistaken: Bargh et al's prime set (like the materials used in many priming studies) utilizes a high proportion of English adjectives, however the frequencies and distribution of adjectives varies considerably across languages (Dye, Milin, Futrell, & Ramscar, 2015), and adjectives appear to play very different functional roles in discourse depending on the language in question (Ramscar et al, 2010; Furtell & Ramscar, 2011; Arnon & Ramscar, 2012; Ramscar 2013; Dye et al, 2015).

To explain why this kind of linguistic variance is particularly relevant to priming research, it is important to note that in speech, and language processing more generally, sequentially encountered acoustic and lexical regularities (phones, words, etc.) serve as cues to upcoming aspects of linguistic signals (MacDonald, Pearlmutter, & Seidenberg, 1994; MacDonald & Seidenberg, 2006; Chang, Dell, & Bock, 2006; Levy, 2008; Ramscar et al., 2010; Jaeger, 2010; Balling & Baayen, 2012; MacDonald, 2013; Ramscar & Baayen, 2013; Ramscar 2013). Not only is speaker behavior sensitive to, and influenced by, contextual information – i.e., the degree to which material is cues, or primed – but a large and ever growing body of empirical evidence indicates that many of the specific distributional properties of languages help to regulate the moment-to-moment uncertainty associated with speech signals. For example, processing nouns involves more uncertainty than other parts of speech (because nouns are more diverse and less predictable than other parts of speech), and languages appear to have evolved a variety of different features to help regulate this uncertainty: For example, in many languages, noun class (or gender) makes nouns more predictable, and easier to process in context (Lew-Williams, & Fernald, 2010; Grüter, Lew-Williams, & Fernald, 2012; Arnon & Ramscar, 2012).

To give an example of what this means in practice, Dye et al (2015) note that in German, while most of the words for drinks have a feminine or masculine gender, three very frequent drink words (*beer, water* and the word *drink* itself) are neuter. Because *beer*, *water* and *drink* are highly frequent drink words, this means that in discourse in which words relating to drinks may be expected to occur, gendered articles provide important discriminative information about which drink words actually will occur (Dye et al, 2015; Arnon & Ramscar, 2012).

English, of course, lacks gender, however corpus studies indicate that, in informational terms, adjectives such as *nice* and *cold* can perform broadly similar functions in English. Although the food and beverage contents of a refrigerator are all cold, upon hearing the phrase *"would you like a nice cold . . . '* the expectations of a native British-English speaker in a refrigerator context will likely come to resemble those of a native German speaker on hearing a neuter article in a similar context (in this context, *nice cold* makes it more likely that *beer* and *drink* will occur, whereas *tonic water* and *wine* become less likely; Dye et al, 2015).

That is, it appears that as it evolved over time and lost its noun-classes, English developed a more informative distribution of adjectives; And critically for current purposes, the distribution of English adjectives appears to be quite different to that found in neighboring gendered languages such as German, French and Spanish (for a detailed illustration of the differences found in the distribution of English and Spanish color adjectives, see Ramscar, 2014a).

It is important that to note at this point that our explanation for the differences in the distribution of adjectives in French and English is, by necessity, a theoretical claim. Pinning down the exact reasons for diachronic change in language is a difficult, perhaps ultimately impossible, task. However, one advantage that theories bring to the scientific endeavor is that they generate predictions that allow hypotheses to be falsified, or refined, and even for credibility to be added to them (see e.g., Popper, 1959). In this case, the theory articulated by Dye et al (2015; see also Ramscar, Milin, Furtell & Dye, in prep) makes clear predictions about the distribution of French and English adjectives:

1. Adjectives will be far more frequent in English than in French,

2. English adjectives will occur in pronominal positions – i.e. before nouns – far more often than French adjectives (and because of this, English adjectives will prime nouns more than French adjectives).

If they are accurate, these predictions would lead us to expect the very different patterns of priming from the Bargh et al. English prime set and the Doyen et al. French prime set. So to test them, we analyzed two matched corpora: UKWaC, (2.25 billion English words), and FRWac (1.6 billion French words; Baroni, Bernadini, Ferraresi & Zanchetta, 2009)[4] . Each corpus was processed using TreeTagger (Schmid, 1994) to annotate each word with a part-of-speech tag, prior to lexical and part of speech counts being taken.

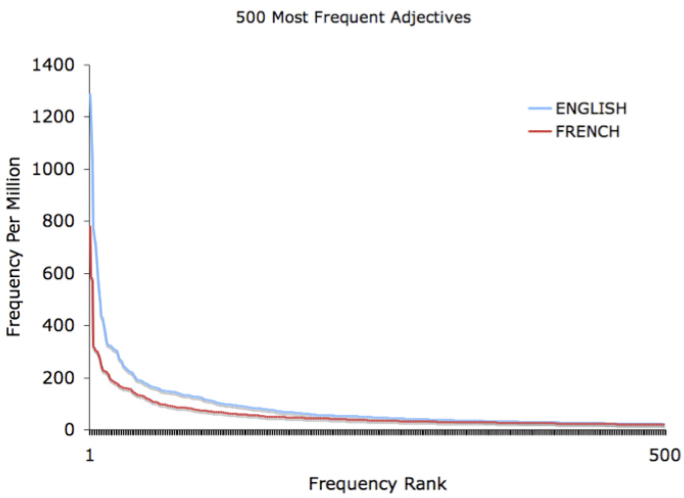

Figure 9: Frequency distribution of the 500 most common adjectives in the UKWaC and FRWaC corpuses (Baroni, Bernadini, Ferraresi & Zanchetta, 2009).

[4]Downloaded from http://wacky.sslmit.unibo.it in December, 2012.

The first of these counts is plotted in Figure 9. As predicted, it shows that adjectives are encountered far more frequently in the English than in the French corpus. In particular, each English adjective is far more frequent than the French adjective of equivalent frequency rank (in a comparison of the 5000 most frequent adjectives in UKWaC and FRWaC, English Mean=89/million, Median=12.5/million; French Mean=59, Median=10; log(frequency) paired $t(9998)=7.1$, $p<0.000001$). Consistent with our predictions about the overall distributional patterns in each language, an analysis of the items in the two prime sets revealed that average frequency of the Bargh et al. items in the English corpus was far higher than that observed in the French corpus for the set of translations (or replacements) used in Doyen et al.'s replication attempt (English, M=33/million, French M=9/million; log(frequency) $t(17)=5.628$, $p=0.0001$, two tailed; Figure 10).

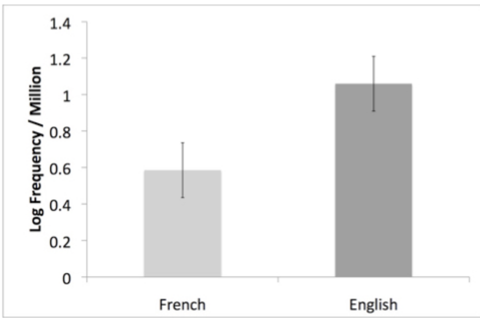

Figure 10: Average log frequencies of the prime words used by Bargh et al. (1996) and Doyen et al. (2012).

We noted above that the majority of the items employed by Bargh et al. and Doyen et al. typically appear as adjectives in speech, however, a few – e.g., Florida – are more commonly encountered as nouns. Interestingly, almost all of the items that are common to the two sets of primes employed by Bargh et al. and Doyen et al. are adjectives (this maybe because the non-adjective items such as Bingo and Florida are particularly culturally specific). To examine the properties of just those items that can be directly compared, we conducted a further comparison of the 21 the mutually translatable primes common to both prime sets. This revealed that whereas the translatable items have a mean frequency of 30/million in English, the frequency of the corresponding items in French is only 8/million (log(frequency) paired $t(20)=4.228$, $p=0.0004$, two tailed).

Finally, to test our prediction about the distribution of these adjectives, and hence their contribution to the overall priming potential of adjectives in the two languages, we compiled contextual co-occurrence counts to establish the degree to which the nouns in each language (proper nouns were excluded) are actually cued (i.e. primed) by adjectives in discourse. All the nouns and adjectives in the corpora were analyzed, including comparatives and superlatives, and the methodology employed, which placed each target noun in the center of a five word window and took co-occurrence frequencies from the two words preceding and following it, restricted the counts to only those contexts in which nouns and adjectives appeared within sentences.

As predicted, this analysis revealed that the distribution of adjectives and nouns differs between the two languages, with 28% of English nouns being preceded by at least one adjective, as opposed to just 10% of French nouns. With regards to the adjectives in the Bargh et al prime set and their translations, the items in the English prime set occurred prior to nouns on average 18 times per million words, whereas the average frequency with which Doyen at al's translations occurred in this position in French is just 3 / million words (log(frequency) paired $t(20)=6.582$, $p<0.0001$, two tailed).

In other words, when it comes to their experience of encountering the items in the prime sets in contexts where they actually served as primes to nouns (e.g., man, woman), we can expect that the subjects in Bargh et al.'s study will have had something of the order of six times more experience than Doyen et al.'s subjects. Even before we consider the effects of linguistic burstiness on the distribution of these experiences across the respective subject populations, we can thus be confident that the experience that serves to shape the expectations of English subjects in contexts in which these words occur will differ considerably from that of French subjects (Figure 11).

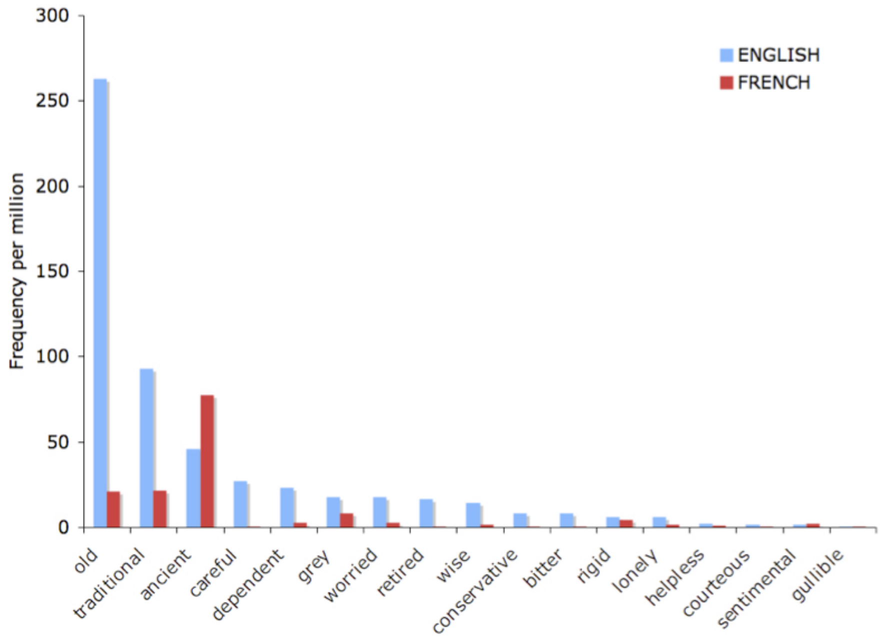

Figure 11: Frequency counts (per million) for the pre-nominal occurrences of the Bargh et al. (1996) and Doyen et al. (2012) prime items that are mutually translatable. The counts were taken from the UKWac and FRWaC corpora (Baroni et al., 2009). As can be clearly seen, the frequency with which the sets of adjectives prime nouns such as *man, woman, lady,* etc. in each of the two languages is very different.

In summary, the English items employed by Bargh et al. in their original study are far more frequent than the French items employed by Doyen et al. in their attempt to replicate this study, and the English items also occur far more frequently as cues to nouns than do their French equivalents. Given the very large differences between the two sets of items that our analysis has revealed, it would in fact be surprising in the extreme if they were to prime behavior in exactly the same way. Given this, it is thus far from surprising that the behavior Bargh et al. reported after priming their English-speaking subjects with their English prime set failed to materialize in French-speaking subjects primed with French items (Doyen et al., 2012). The different properties of these words in the two languages offers good reasons to expect that a valid priming effect based on English items will fail to replicate with French translations of them. Indeed, the linguistic differences we describe suggest that it is effectively impossible for anyone to directly replicate Bargh et al. (1996) Experiment 2 in French.[5] Given that priming effects clearly are a function of the effect that exposure to a prime (or set of primes) has on the expectations of subjects, it follows that unless the factors that shape the learning of those expectations (such as the frequency and distribution of lexical cues) are controlled for (along the lines outlined here), failed replication attempts in other languages that are conducted using naive translations of materials in other languages can only be seen as scientifically uninformative, and something of a waste of time and resources.

However, although our analyses indicate that other 'failed replication' studies which have been conducted in translation (see e.g., Gámez, Díaz, & Marrero, 2011; Aveyard, 2014) cast as little light on the effects that they studied as Doyen et al.'s (2012) study does, we are hopeful that the methods we have utilized in these analyses can provide a blueprint for the way that priming materials might be analyzed in order to establish whether a priming effect translates, and if so, whether it can be replicated in a different language.

3. Why priming studies cannot be expected to replicate across the lifespan

Thus far we have shown that once the lexical and cultural properties of prime sets are considered, there is reason to believe that many priming results that rely on lexical (or other symbolic) stimuli for their efficacy will be unlikely to replicate over time or between languages. Since the point of many scientific studies in psychology is to identify or establish generalizations about "how the mind works" (Bargh, 2014), we next consider the question of whether priming results can be expected to replicate between different age groups in the same social and linguistic population, or – notwithstanding the problems raised above – even within the same individual across the lifespan.

The answer, we shall suggest, is "no" and the reason for this answer lies in the learning mechanisms that we assume give rise to priming effects in the first place. Crucially, although there is widespread acceptance in the literature that priming is driven by the activation of 'learned associations,' characterizations of what learning actually involves in

[5] Although the majority of classic social priming studies were conducted in English, many of the more notorious fraud cases in this literature involve research carried out in other languages (Yong, 2012). It is likely that the particular function of English adjectives makes the language a particularly fertile one in which to conduct priming studies, and that other superficially similar languages offer less fertile grounds in which to find similar effects. This in turn raises the question of whether these subtle differences have played a role in leading astray investigators conducting priming studies in these less amenable languages.

the priming literature tend to be both vague and inaccurate. For example, Dijksterhuis & Bargh (2001) characterize priming as, "the activation of a perceptual representation, [which] has a direct effect on social behavior. Perceptual inputs are translated automatically into corresponding behavioral outputs"

However, although Dijksterhuis & Bargh clearly assume that the links between 'perceptual representation[s]' and 'social behavior' are formed by learning, they go on to reject the idea that this results from any known learning process:

"The . . . behaviorists' thesis that responses follow directly from perceived stimuli, or S-R bonds, also holds that perception directly leads to action (e.g., Skinner, 1938; Watson, 1913). However, the mechanism proposed is quite different, because these responses are not imitations of the perceived event but are stamped in responses to stimuli based on one's past reinforcement history. The behaviorist . . . would argue, as we do here, that behavioral tendencies are put into motion directly by perceptual activity, but unlike the present theme, they also argue (more or less) that these tendencies are learned responses over time based on one's history of reward and punishment with those stimuli . . . it is not necessary that the behavioral response be stamped in as a habit through reinforcement" (Dijksterhuis & Bargh, 2001)

What is striking about this account – apart from the questions it raises about the relationship between prime words and experience – is the very outdated account of associative learning that is described by the authors.

It is clearly a fact that any relationship between the word 'old' and any priming properties it may have must be learned, simply because old is an English cultural construct (Ramscar & Port, 2015), a word that is acquired by almost anyone who grows up in an English-speaking home, regardless of their particular geographical and ethnic origins. Importantly, the basic learning mechanisms that Dijksterhuis & Bargh (2001) dismiss have been shown to be very helpful in explaining – and even formally predicting – just how it is that children and adults go about learning these words and how they subsequently go about processing and using them (Arnon & Ramscar, 2012; Baayen & Ramscar, 2015, Baayen et al, in press, 2012; Baayen, 2012; Baayen, Milin, Filipovic Durdevic, Hendrix, & Marelli, 2011; Mulder, Dijkstra, Schreuder, & Baayen, 2014; Ramscar et al, 2014, 2013abcd; Ramscar, 2013; Ramscar, Dye, Popick & O'Donnell-McCarthy, 2011; Ramscar et al, 2010b).

Further, Dijksterhuis & Bargh's suggestion that reinforcement learning is a process driven by rewards and punishments has long been shown to be largely irrelevant to our understanding of 'associative learning,' and as is clear from the relevant literature, students of learning abandoned these notions over half a century ago (for reviews see Rescorla, 1988; Ellis, 2006; Ramscar et al., 2010; Ramscar et al., 2013b). [6] The widespread yet

[6] This naïve view of learning is hardly unique to Dijksterhuis & Bargh: Some 25 years ago, Rescorla (1988) lamented the woefully outdated understanding most psychologists have of learning theory, and the poor teaching of learning in psychology departments, yet it is clear that little has changed since (see, e.g. Bransford et al, 2004).

mistaken belief that learning is driven by punishment and reward most likely persists in the literature because punishment and reward schedules are widely employed in animal learning experiments in order to make the behavior of animals interpretable (i.e., not because researchers believe that they are necessary for learning). Learning itself has long been understood to be driven by information in the environment: Animals do not dumbly 'associate' stimuli with punishments and rewards, but rather, as a vast array of findings has shown, animals' minds are engaged in trying to predict events in the world, and learning acts to discriminate the aspects of the environment that best support these predictions (Rescorla, 1988; Ramscar et al, 2010b, 2013a,b; Baayen, Shaoul, Willits, & Ramscar, in press).

This research has revealed that discriminative learning processes are driven by error: The value of cues that prime expectations that are violated by experience get downgraded (discriminated against) by this process, which in turn positively discriminates in favor of cues that prime more reliable expectations. Learning is thus an 'embodied' process in which the totality of sensory and experiential information available to learners competes dynamically for relevance, incrementally helping to produce a 'mental model' of the predictive structure of the environment (Ramscar et al, 2010b, 2013b).

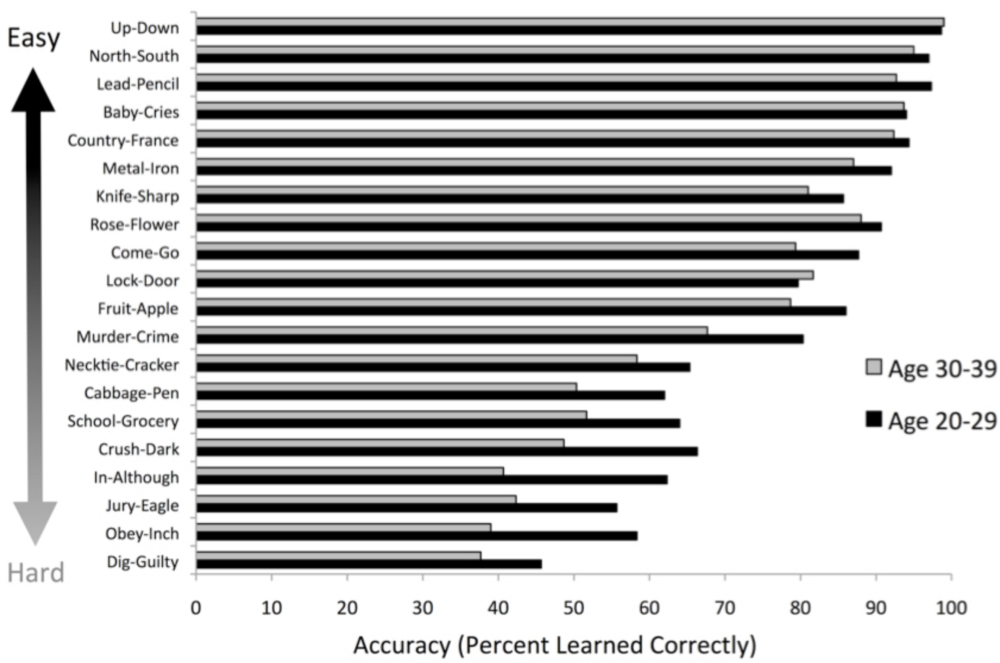

Figure 12: Average by-item performance for 400 adults aged 20–29 and 30–39 (50 males & 50 females per group) on forms 1 & 2 of the WMS-PAL subtest (desRosiers & Ivison, 1988). As can be seen, whereas there is little change in the way 'easy' items prime their targets between the twenty-something and thirty-something groups, the amount of priming in the 'hard' pairings changes far more with increased experience.

The relevance of learning processes to priming can be illustrated by considering to a simple, much-studied associative priming task, paired associate learning (PAL): in this task, subjects learn pairings between word primes (e.g., baby; jury) and responses (cries; eagle), and are then tested on their ability to produce the correct response: e.g., respond eagle after being presented with the prime word jury. As can be seen in Figure 12, the degree to which words serve to successfully prime their targets in these tests declines as subjects' age increases, such that the average change in Figure 12 might seem to provide clear evidence that adult associative learning capacity (and hence prime word efficacy) declines considerably between age 29 (mean PAL Accuracy 78%) and 39 (70%; $t(19)=5.286$, $p<0.0001$). This change in PAL performance has traditionally been attributed to "cognitive decline" (Salthouse, 2009; Singh-Manoux et al., 2012; see also Nolan & Johnson, 1967), and has commonly been supposed to reveal declines in associative learning capacity.

However, Ramscar et al (2013b, 2014) have shown that these changes can also be explained by reflecting on the way discrimination learning can be expected to reshape priming potentials across sets of items over time. To understand why learning might affect priming in PAL tests in the way depicted in Figure 12, it may help to consider that to a Mandarin monolingual, the word pairs in Figure 12 are likely to be all equally easy or hard to learn. However, if our Mandarin monolingual was to learn English, she might learn that some pairs share similarities of meaning – e.g., up - down – and this may increase the degree to which up primes down, making learning this pairing easier.

What the data plotted in Figure 12 make clear is that the actual differences between older and younger subjects' PAL performance are greatest for the hard items, not the easy ones. This raises the question – especially given Dijksterhuis & Bargh's portrayal of learning above – of how it is that experience makes priming in the 'hard' PAL pairs decrease. The answer lies in the way that evidence about error (the non-occurrence of expected events) accumulates as people learn more about the predictive structure of the world from experience (see Ramscar et al, 2010b, 2013b for reviews). As our Mandarin speaker encounters more and more Prime words (w1) without Target words (w2) in the course of her English studies, the pairing of w1 - w2 will become harder to learn, because each occurrence of w1 without w2 will reduce the likelihood that w1 is informative about w2. She will effectively learn that w1 does not prime w2 (Ramscar et al, 2010b, Ramscar et al, 2013c). Similarly, because error-driven learning serves to systematically reduce uncertainty (well-anticipated events promote less learning than unanticipated events, Ramscar et al, 2010b) infrequently encountered Target (w2) items will support better learning of the relationship between w1 and w2 than items that are frequently encountered, simply because frequent words will, by their very nature, become more predictable over time (Kamin, 1969). As a corollary, w1-w2 pairs will tend to become harder to learn when w1 and w2 occur independently at high rates, or where the independent frequency of w2 is far greater than w1, and easier to learn when w1 and w2 co-occur at high rates (Rescorla & Wagner, 1972).

It follows that if we assume that associative priming is influenced by associative learning (which, computationally, is a discriminative process, see Ramscar et al., 2010b), we can thus estimate the degree to which w1 can be expected to prime w2, in the limit, based on (1) the frequency of w1,(2) its relationship to the predictability of w2 and (3) the actual rate at which w1 and w2 co-occur. When we conducted this analysis on the data in Figure 12 using the log of these lexical paramaters (drawn from large corpora, Ramscar et al, 2013c, 2014)[7] as the input to simple linear regression models, they account for 85% of the by-item variance in the PAL performance in both the 20-29 ($(F(3,16) = 31.01, r > .9, p < .01)$ and 30–39 ($F(3,16) = 30(64, r > .9, p < .01)$ age groups. Further, as we would expect given this analysis, the word frequency parameters that determine the independence of w1 and w2 were associated with lower scores, while the w1-w2 association-rate parameter was associated with higher scores (all six tests, $p < .01$), while the regression beta weights for the predictors in the age 30–39 group were more extreme (farther from zero) than those for the 20–29 group.

It is worth noting here that despite Dijksterhuis & Bargh's claim that conditioning is insufficient to explain social-priming, these results reveal that it does a remarkably good job of accounting for the variation in the priming between individual items in PAL experiments. Indeed, given that the burstiness of language necessarily places an upper bound on the amount of behavioral variation that we might expect to account for in terms of average performance across subjects, it isn't clear that one can expect to account for much more priming variance that our averaged parameters actually do here (for a replication of this analysis using different corpus data, see Ramscar, 2014b).

To return to our discussion of learning and priming, these results strongly indicate that priming results found in one age range ought not be expected to necessarily replicate between different age groups within the same social and linguistic population, or even across the course of the same adult's lifespan, simply because the degree to which words prime other words (or "concepts," etc.) can and will change dramatically with experience (and given that the mechanisms that drive these changes are not specific to language, this point is also likely to apply to other kinds of priming stimuli).

Further, insofar as changes in PAL performance in adulthood suggest that older adults both know more words (Keuleers et al, 2014), and their knowledge of how to use these words is better discriminated (Ramscar et al, 2013), it is likely that their behavior will be increasingly likely to be influenced only by more specific primes, and less likely to be influenced by less-specific primes, a suggestion that is supported by studies of PAL performance with bi- and monolinguals (Ramscar, Sun, Hendrix & Baayen, 2015): whereas intuition might suggest that priming between associates would be stronger in one's mother tongue, older Chinese-German bilinguals' PAL performance is both higher in their second language than their first, and better than that of age-matched native German speakers.

Given that many results cannot and should not be expected to replicate, where does this leave priming research?

In the Cratylus, Plato examines Heraclitus' claim that everything is in a state of continuous flux: If everything is constantly changing, he asks, how can one ever know

[7] W1 and w2 word frequencies were taken from the COHA 1970–1980 sample (Davies, 2012) and log transformed; the relationship between w2 and w1 predictability was calculated as (log(w2 frequency) / log(w1 frequency)); w1-w2 co-occurrence rates were taken from Google and log transformed.

anything?

"how can that be a real thing which is never in the same state? ...for at the moment that the observer approaches, then they become other ...so that you cannot get any further in knowing their nature...Nor can we reasonably say ...that there is knowledge at all ...for knowledge too cannot continue to be knowledge unless continuing always to abide and exist. ...but if that which knows and that which is known exist ever ...then I do not think that they can resemble a process or flux...no man of sense will like to put himself or the education of his mind in the power of names: neither will he so far trust names or the givers of names as to be confident in any knowledge which condemns himself and other existences to an unhealthy state of unreality" Plato, Cratylus (paragraph 440).

Philosophers have tended to address Plato's questions by focusing on the verb to know, and have spent much time pondering the nature of epistemic certainty. In contrast, the natural sciences have tended to eschew epistemology, and have focused instead on determining the reliability of theoretical conjectures about a universe that really does appear to be in flux. Rather than seeking to establish criteria for certainty, the natural sciences have sought to manage and minimize uncertainty (such that learning and the pursuit of science share obvious common attributes; Rescorla, 1988).

Scientists (and, critically, statisticians) accept that all models and theories are wrong, to a degree (Box & Draper, 1987), and scientific understanding has flourished – and flourishes – through an iterative process of establishing those circumstances in which models and theories succeed in providing coverage, as well as those in which they fail to (see also Cesario, 2014). Replication — the independent confirmation of results from one study in another — offers a scientific gold standard in this regard, because it helps to empirically delineate the phenomena that models and theories must account for (Jasny, Chin, Chong & Vignieri, 2011; Cesario, 2014; Simons, 2014).

Replicability is, however, an inherently contingent matter: Experimental effects can only ever be expected to replicate when the natural phenomena that are being studied by scientists are sufficiently invariant (in terms of time and context) to allow them to. Indeed, empirically, it seems clear that Heraclitus' flux is to some degree a fact of scientific life: Studies have shown that even materials as seemingly solid as metals have properties that are more akin to gases than our common intuitions are wont suppose (see e.g., Scerri, 1991). However, most structural aspects of most metals (along with much of the rest of the physical universe) are sufficiently invariant for most intents and purposes to ensure that exact replications of studies of produce highly similar set of results (probabilistically, at least) in most circumstances.

Similarly, while our understanding of the biological world actively embraces the Heraclitean flux (Darwin, 2009), biological changes usually happen at a rate that allow for replication. And because our understanding of biology has grown beyond a mere catalog of events, that even when studies ?fail to replicate? exactly on a trial by trial basis (because the organisms being studied evolve in the course of study), the changing nature of these results actually serves to illuminate theories about the mechanisms of change, e.g., Blount, Borland, & Lenski, 2008).

The scientific process that emerges from the replicability of biological results is evident in the development of our understanding of learning. A reliable body of replicable results established a set of empirical phenomena that provide both a spur and a target for the development of models to predict and explain the empirical data (e.g., Rescorla & Wagner, 1972; Pearce & Hall, 1980; Danks, 2003). While none of these models account for all the relevant empirical phenomena (all models are wrong in the limit), their commonalities – especially in the use of discrepancies between expected and experienced outcomes to drive learning – led in turn to the identification of the brain structures that appear to compute the results of these discrepancies (Waelti, Dickinson, & Schultz, 2001; Schultz, Dayan, & Montague, 1997; Schultz & Dickinson, 2000; Schultz, 2006; 2010). The existence of a replicable body of evidence allowed for the development of models, the testing of model predictions, and the development of a body of congruent findings and models from studies employing different methods that, taken together, yields a good deal of useful scientific and practical insight.

Although priming effects clearly have a biological component – it is highly likely that they are a product of many of the same predictive brain processes involved in learning – many, if not all, social priming effects also rely for their efficacy on cultural and experiential factors that are clearly not invariant over time. Rather, as we have shown, the dynamics of learning and the dynamics of cultural change mean that perfectly valid social priming results cannot and should not be expected to replicate in the same way that valid results from physical and biological studies tend to do. In other words, specific priming results cannot be expected to replicate because their social and cultural components are not invariant aspects of the natural world, and it is the relative invariance of the natural world that enables replication to be the cornerstone of cumulative natural science.

The problem that priming research has to face is not that the field doesn?t replicate its findings often enough, but rather that, as they are currently employed, the approaches that some researchers in psychology have borrowed from the natural sciences may not be appropriate to the phenomena they study at the level that actually study them. That is, while replication is a critical source of information about the reliability of research findings, it does not follow from this that the validity of findings is best evaluated by either direct replication (see e.g., Nosek et al, 2012; Asendorpf, et al., 2013), or by conceptual replications, in which different materials are employed to study the 'same' priming effect (Ritchie, Wiseman, & French, 2012; Stroebe & Strack, 2014).

Rather, we would suggest that a great many findings in the psychological literature would be more usefully and profitably evaluated by the kinds of analyses we present here than by going through the motions of trying to replicate studies that are in fact irreplicable. Although the intentions behind recent moves towards conducting direct replications of psychological findings en mass are laudable (see e.g., Nosek et al, 2012; 2015), the degree to which this strategy can evaluate the validity of socially and culturally embedded findings absent the kind of broader analyses presented above is severely limited, simply because it is inevitable that the dynamics of change in social and cultural phenomena will cause many perfectly valid social priming findings published today to fail to replicate in 2062 regardless of the fact that they may replicate in the shorter term. Accordingly, the adoption of objective methods for analyzing and quantifying these things can only help to put this research on a firmer footing, by helping to explain both why it is that priming does work, and in accounting for why many specific effects may be transitory. In the long run, we believe that methods of analysis of the kind we present here have more to offer than helping make scientific sense of the phenomena addressed in social priming research: Recent training practices have left generations of psychologists with a very poor understanding of learning (Rescorla, 1988), and language and information theory have faired no better (Rescorla, 1988; Ramscar, Hendrix & Baayen, 2012; Ramscar et al, 2010, 2013b). One important reason for this state of a affairs is that experimental psychologists, having largely accepted some speculative, epistemological claims about the ability of specific theories of learning to explain specific theories of language in the middle of the last century (Virués-Ortega, 2006), have largely ignored learning (and language) ever since. Indeed, we would suggest that the mismatch between what psychologists know and what they ought to know about language and learning that we highlighted in our analyses of PAL learning above (see also Ramscar et al 2013c, 2014) is a far larger problem than the methodological concerns about priming which currently preoccupy the field is. Not only will addressing this problem do far more for our understanding of priming than any number of attempts to 'replicate' culturally transient findings, but given the essentially linguistic nature of homo sapiens, it is also likely that doing so will prove to be an essential prerequisite to establishing a firmer scientific footing for the psychological sciences; and for properly addressing the more important psychological questions, such as, for example, understanding what effects aging actually has on our minds and brains.

References

Arnon, I. & Ramscar, M. (2012) Granularity and the acquisition of grammatical gender: How order of acquisition affects what gets learned. *Cognition*, 122(3), 292-305

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., ... & Wicherts, J. M. (2013). Replication is more than hitting the lottery twice. *European Journal of Personality*, 27(2). p.138-138

Aveyard M.E. (2014) A Call to Honesty: Extending Religious Priming of Moral Behavior to Middle Eastern Muslims. *PLoS ONE* 9(7): e99447. doi:10.1371/ journal.pone.0099447

Baayen, R. H., Shaoul, C., Willits, J. and Ramscar, M. (in press). Comprehension without segmentation: A proof of concept with naive discrimination learning. *Language, Cognition, and Neuroscience.*

Baayen, R. H. and Ramscar, M. (2015). Abstraction, storage and naive discriminative learning. In Dabrowska, E. and Divjak, D. (eds) , *Handbook of Cognitive Linguistics.*, 99-120. Berlin: De Gruyter Mouton

Baayen, R.H., Hendrix, P., & Ramscar, M. (2013) Sidestepping the combinatorial explosion: An explanation of n-gram frequency effects based on naive discriminative learning.*Language & Speech*

Baayen, R. H. (2012) Learning from the Bible: Computational modeling of the costs of letter transpositions and letter exchanges in reading Classical Hebrew and Modern English. *Lingue & Linguaggio* 2, 123-146

Baayen, R. H., Milin, P., Filipovic Durdevic, D., Hendrix, P. and Marelli, M. (2011), An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review* 118, 438-482.

Baayen. R.H. (2001) Word Frequency Distributions. Kluwer Academic Publishers.

Balling, L. and Baayen, R.H. (2012) Probability and surprisal in auditory comprehension of morphologically complex words. *Cognition* 125, 80-106

Bargh JA, Chen M, Burrows L (1996) Automaticity of social behavior: direct effects of trait construct and stereotype-activation on action. *Journal of Personality and Social Psychology* 71: 230?244.

Bargh, J. A. (2012a, March 5). *Nothing in their heads* [Web log post]. Retrieved from Retrieved June 29, 2015, from http://web.archive.org/web/20120309182006/http://www.psychologytoday.
natural-unconscious/201203/nothing-in-their-heads?

Bargh, J. A. (2012b, March 23). *Angry birds: Nothing in their heads,* Round 2 [Web log

post]. Retrieved June 29, 2015, from http://web.archive.org/web/20120325213848/http://www.psychol
natural-unconscious/201203/angry-birds

Bargh, J. A. (2012c, May 11). *Priming Effects Replicate Just Fine, Thanks.* Psychology Today. Retrieved June 29, 2015, from http://www.psychologytoday.com/blog/the-natural-unconscious/201205/priming-effects-replicate-just-fine-thanks

Baroni, M. Bernardini, S. Ferraresi, A and Zanchetta, E (2009) The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation* 43(3): 209-226

Blount, Z. D., Borland, C. Z., & Lenski, R. E. (2008). Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli. *Proceedings of the National Academy of Sciences,* 105(23), 7899-7906.

Boroditsky, L. & Ramscar, M. (2002). The roles of body and mind in abstract thought. *Psychological Science*, 13(2), 185-188

Box, G.E.P. & Draper, N.R. (1987).*Empirical Model-Building and Response Surfaces,* Wiley. ISBN 0-471-81033-9.

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science*, 9, 40?48.

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psycological Review*, 113, 234?272.

Darwin, C. (2009). *The origin of species by means of natural selection: or, the preservation of favored races in the struggle for life.* W. F. Bynum (Ed.). AL Burt. Davies, M (2012) Expanding Horizons in Historical Linguistics with the 400 million word Corpus of Historical American English. *Corpora* 7: 121-57.

Des Rosiers G, Ivison D. (1986) Paired-associate learning: normative data for differences between high and low associate word pairs.*J Clin Exp Neuropsychol*; 8: 637?42.

Dijksterhuis, A., & Bargh, J. A. (2001). The perception-behavior expressway: Automatic effects of social perception on social behavior. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 33, pp. 1?40). San Diego: Academic Press.

Doyen, S., Klein, O., Pichon, C.L., & Cleeremans, A. (2012). Behavioral priming: it's all in the mind, but whose mind? *PloS One,* 7 (1) PMID: 22279526

Dye, M., Milin, P., Futrell, R., & Ramscar, M. (2015). A functional theory of gender paradigms. In F. Kiefer, J.P. Blevins, & H. Bartos (Eds.) Brill: Leiden.

Ellis, N. C. (2006). Selective Attention and Transfer Phenomena in L2 Acquisition:

Contingency, Cue Competition, Salience, Interference, Overshadowing, Blocking, and Perceptual Learning. *Applied Linguistics*, 27 (2), 164-194

Firth, J.R. (1957). A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*, pp. 1-32. Oxford: Philological Society. Reprinted in F.R. Palmer (ed.), Selected Papers of J.R. Firth 1952-1959, London: Longman (1968).

Forster, K., & Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10 (4), 680-698

Futrell, R & Ramscar M (2011) German Grammatical Gender Manages Nominal Entropy. Presentation at Information-theoretic Approaches to Linguistics, 2011 *LSA Linguistic Institute,* Ohio State University

Gámez, E., Díaz, J. M. & Marrero, H. (2011) The uncertain universality of the Macbeth effect with a Spanish sample. *Spanish Journal of Psychology* 14:156?62.

Grüter, T., Lew-Williams, C., & Fernald, A. (2012). Grammatical gender in L2: A production or a real-time processing problem?.*Second Language Research*, 28(2), 191-215.

Hutchison, K. A., Balota, D. A., Cortese, M. J. and Watson J. M. (2008) Predicting semantic priming at the item level. *The Quarterly Journal of Experimental Psychology* Vol. 61, Iss. 7

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density.*Cognitive Psychology*, 61(1), 23?62

Jasny, B. R., Chin, G., Chong, L., & Vignieri, S. (2011). Again, and again, and again?. *Science*, 334(6060), 1225-1225.

Johns, B. T., & Jones, M. N. (2010). Evaluating the random representation assumption of lexical semantics in cognitive models. *Psychonomic Bulletin & Review,* 17, 662-672

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, 55, 534-552

Jones, M. N., & Mewhort,D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, 114, 1-37.

Kahneman, D. (2012). A proposal to deal with questions about priming effects. *Nature.* doi, 10.

Kamin L.J. Predictability, surprise, attention, and conditioning. In: Campbell B, Church R (eds). *Punishment and Aversive Behaviour.* Appleton-Century-Crofts: New York, (1969).

Lew-Williams, C., & Fernald, A. (2010). Real-time processing of gender-marked articles by native and non-native Spanish speakers. *Journal of Memory and Language*, 63, 447-464.

MacDonald, M. C., Pearlmutter, N., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, 101, 676?703.

MacDonald, M. C., & Seidenberg, M. S. (2006). Constraint satisfaction accounts of lexical and sentence comprehension. In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics*, 2nd edition (pp. 581?611). London: Elsevier.

Matlock, T, Ramscar, M., & Boroditsky, L. (2005). The experiential link between spatial and temporal language. *Cognitive Science*, 29, 655-664.

Meyer, D.E., & Schvaneveldt, R.W. (1971). Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *Journal of Experimental Psychology,* 90 (2), 227-34 PMID: 5134329 Milin, P., Ramscar, M., Cho, K., Baayen, R.H., & Feldman, L. (2015) *Cornering segmentation: the perspective from discrimination learning.* Manuscript under review.

Mulder, K., Dijkstra, T, Schreuder, R. and Baayen, R. H. (2014). Effects of primary and secondary morphological family size in monolingual and bilingual word processing. *Journal of Memory and Language*, 72, 59-84

Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner & G. W. Humphreys (Eds.), *Basic processes in reading: Visual word recognition* (pp. 264?336). Hillsdale, NJ: Lawrence Erlbaum Associates

Nolan, W.F. & Johnson, G.C. (1967) *Logan's Run*, New York, NY: Dial Press

Nosek, B.A., & The Open Science Collaboration (2012) An Open, Large-Scale, Collaborative Effort to Estimate the Reproducibility of Psychological Science, *Perspectives on Psychological Science* 2012 7: 657-660

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... & Yarkoni, T. (2015). Promoting an open research culture.*Science,* 348(6242), 1422-1425.

Pashler, H., Coburn, N., and Harris, C. (2012). Priming of social distance? Failure to replicate effects on social and food judgements. *PloS ONE*, 7(8):, e42510 Pashler, H., Harris, C., & Coburn, N.. *Elderly-Related Words Prime Slow Walking* . (2011, September 15). Retrieved 07:52, December 12, 2012 from http://www.PsychFileDrawer.org/replication.php?attem

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli.*Psychological Review*, 87, 532?552.

Plato (1953) tr. Jowett, B. *Cratylus*, in The Dialogues of Plato, Oxford: OUP.

Popper, K. R. (1959). The logic of scientific discovery. London: Hutchinson.

Ramscar, M., Sun, J., Hendrix, P & Baayen, R.H. (2015) The mismeasurement of mind: What paired associate learning tasks reveal about cognitive performance across the lifespan. Manuscript, Seminar für Sprachwissesschaft, University of Tübingen.

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P. & Baayen, R.H. (2014) The myth of cognitive decline: Non-linear dynamics of lifelong learning. *Topics in Cognitive Science*, 6, 5-42.

Ramscar, M. (2013) Suffixing, prefixing, and the functional order of regularities in meaningful strings. *Psihologija*, 46(4), 377-396

Ramscar, M., Dye, M., & Klein, J. (2013a). Children value informativity over logic in word learning. *Psychological science*, 24, 1017?1023

Ramscar, M., Dye, M., & McCauley, S. M. (2013b). Error and expectation in language learning: The curious absence of mouses in adult speech. *Language*, 89, 760?793.

Ramscar, M., Hendrix, P., Love, B., & Baayen, R. (2013c). Learning is not decline: The mental lexicon as a window into cognition across the lifespan. *The Mental Lexicon,* 8, 450?481

Ramscar, M., Dye, M., Gustafson, J.W., & Klein, J. (2013d) Dual routes to cognitive flexibility: Learning and response conflict resolution in the dimensional change card sort task. *Child Development*, 84(4), 1308-23.

Ramscar, M. (2014a, June 23). *The errors in my answer to Darwin* [Web log post]. Retrieved from Retrieved June 29, 2015, from https://ramscar.wordpress.com/2014/06/23/the-errors-in-my-answer-to-darwin/

Ramscar, M. (2014b, August 7). *Technical Note: Bells and Smells and Replication* [Web log post]. Retrieved from Retrieved June 29, 2015, from https://ramscar.wordpress.com/2014/08/07/tec
note-bells-and-smells-and-replication/

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., & Baayen, H. (2014). The Myth of Cognitive Decline: Non?Linear Dynamics of Lifelong Learning. *Topics in Cognitive Science* 6, 5-42

Ramscar M & Futrell, R (2011) "The Predictive Function of Prenominal Adjectives" Presentation at Information-theoretic Approaches to Linguistics, 2011 *LSA Linguistic Institute,* Ohio State University

Ramscar, M.J.A, Futrell, R, Dye, M, Starr, R, Hendrix, P. & Baayen, R.H. (2012) *Why are names hard to remember? The evolution and decline of universal name grammars.* Manuscript under review.

Ramscar M, Dye M, Popick HM, O'Donnell-McCarthy F (2011) The enigma of number: Why children find the meanings of even small number words hard to learn and how we can help them do better. *PLoS ONE* 6(7): e22501. doi:10.1371/journal.pone.0022501

Ramscar, M., Matlock, T., & Dye, M. (2010a). Running down the clock: The role of expectation in our understanding of time and motion. *Language and Cognitive Processes,* 25(5), 589-615.

Ramscar, M., Milin, P., Furtell, R. & Dye, M. (in prep) Der Pfauenschwanz's Tale: The Evolution Of Noun Classification Manuscript, Seminar für Sprachwissesschaft, University of Tübingen.

Ramscar, M.J.A., D.G. Yarlett, M.W. Dye, K. Denny & K. Thorpe (2010b). The Effects of Feature-Label-Order and their implications for symbolic learning. *Cognitive Science* 34(6), 909-957.

Rescorla R.A. Probability of Shock in the Presence and Absence of CS in Fear Conditioning. (1968). *J. Comp. & Physiol. Psychol.,* 66, 1-5

Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American Psychologist*, 43(3), 151

Rescorla, R.A. & A.R. Wagner. (1972) A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Black & Prokasy (Eds.), *Classical Conditioning II: Current Research and Theory.* (Appleton-Century-Crofts)

Ritchie, S. J., Wiseman, R., & French, C. C. (2012). Replication, replication, replication. The Psychologist, 25 (5), 346-348

Salthouse, T.A. When does age-related cognitive decline begin? (2009). *Neurobiology of Aging* 30, 507-14

Schmid, H (1994) Probabilistic Part-of-Speech Tagging UsingDecision Trees. *Proceedings of International Conference on New Methods in Language Processing,* Manchester, UK.

Schultz, W., & Dickinson, A. (2000). Neural coding of prediction errors. *Annual Review of Neuroscience,* 23, 473-500.

Schultz W. (2006) Behavioral theories and the neurophysiology of reward. *Annual Review of Psychology,* 57, 87-115.

Schultz, W. (2010) Dopamine signals for reward value and risk: basic and recent data. *Behavioral and Brain Functions*, 6, 1-9.

Schultz W, Dayan P., & Montague P.R. (1997) A neural substrate of prediction and reward. *Science,* 275, 1593-1599.

Simons, D.J. (2014) The Value of Direct Replication, *Perspectives on Psychological Science*, 9, 76-80

Singh-Manoux, A. et al. (2012). Timing of onset of cognitive decline: Results from Whitehall II prospective cohort study.*BMJ* 344, d7622

Srull, T. K., & Wyer, R. S. (1979). The role of category accessibility in the interpretation of information about persons: Some determinants and implications. *Journal of Personality and Social Psychology*, 37(10), 1660

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science,* 9(1), 59-71.

Virus-Ortega, J. (2006). The case against BF Skinner 45 years later: An encounter with N. Chomsky. *The Behavior Analyst*, 29(2), 243.

Waelti, P., Dickinson, A., & Schultz, W. (2001) Dopamine responses comply with basic assumptions of formal learning theory. *Nature,* 412, 43-48.

Yong, E. (2012). Replication studies: Bad copy.*Nature,* 485(7398), 298.

Zipf, G. K. (1949). *Human behavior and the principle of least-effort.* Cambridge, MA: Addison-Wesley.