TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 6 (2014) 5–42

ISSN:1756-8757 print / 1756-8765 online
DOI: 10.1111/tops.12078

The Myth of Cognitive Decline: Non-Linear Dynamics of Lifelong Learning

Michael Ramscar, Peter Hendrix, Cyrus Shaoul, Petar Milin, Harald Baayen

Department of Linguistics, Universität Tübingen

Received 19 August 2013; received in revised form 11 October 2013; accepted 24 October 2013

Abstract

As adults age, their performance on many psychometric tests changes systematically, a finding that is widely taken to reveal that cognitive information-processing capacities decline across adulthood. Contrary to this, we suggest that older adults' changing performance reflects memory search demands, which escalate as experience grows. A series of simulations show how the performance patterns observed across adulthood emerge naturally in learning models as they acquire knowledge. The simulations correctly identify greater variation in the cognitive performance of older adults, and successfully predict that older adults will show greater sensitivity to fine-grained differences in the properties of test stimuli than younger adults. Our results indicate that older adults' performance on cognitive tests reflects the predictable consequences of learning on information-processing, and not cognitive decline. We consider the implications of this for our scientific and cultural understanding of aging.

Keywords: Learning; Language; Memory; Psychometric testing

1. The age of Tithonus

More and more people now live longer and longer lives. With the exception of 18 countries the United Nations describes as "outliers," increased life expectancy and declining birth rates are increasing the median age of populations across the globe (Watkins et al., 2005). By 2030, 72 million Americans will be aged 65 or older, a two-fold increase from 2000. The proportion of older adults in the world's population is larger than ever before, and it is growing at an increasing rate.

While it is clear that more people now live longer than ever before in history, it is less obvious that this is a blessing. In Greek mythology, Tithonus was the mortal lover of Eos,

Correspondence should be sent to Michael Ramscar, Universität Tübingen Seminar für Sprachwissenschaft Wilhelmstrasse 19, 72074 Tübingen, Germany. E-mail: michael.ramscar@uni-tuebingen.de

goddess of the dawn. Eos asked Zeus to make Tithonus immortal but failed to mention "eternal youth," dooming Tithonus to an eternity of physical and mental decay. The tithonean account of aging echoes loudly in the literature of the psychological and brain-sciences, which portrays adulthood as a protracted episode in mental decline, in which memories dim, thoughts slow, and problem-solving abilities diminish (Deary et al., 2009; Naveh-Benjamin & Old, 2008), and where researchers seem to compete to set the advent of cognitive decrepitude at an ever younger age (Salthouse, 2009; Singh-Manoux et al., 2012). Thus, although studies indicate that older adults are, on average, happier than younger adults (Charles & Carstensen, 2010), in the light of the foregoing, even this small crumb of comfort might be seen as further evidence of their declining mental prowess.

Because it is believed that cognitive abilities wither over the course of adulthood, population aging is thought to pose a serious threat to the world's economic well-being (Watkins et al., 2005): As the proportion of cognitively impaired adults in the population increases, it is feared they will impose an ever-larger burden on the ever-smaller proportion of society still in full command of its cognitive faculties. Given this uncertain scenario, understanding the way our minds age could be considered the most significant matter that the psychological and brain sciences address.

In what follows, we consider the question of whether one might reasonably expect that performance on any measure of cognitive performance could or should be expected to be age- or, more specifically, *experience*-invariant. We shall suggest that, since the answer to this question is no, many of the assumptions scientists currently make about "cognitive decline" are seriously flawed and, for the most part, formally invalid. We will show that the patterns of response change that are typically taken as evidence for (and measures of) cognitive decline arise out of basic principles of learning and emerge naturally in learning models as they acquire more knowledge. These models, which are supported by a wealth of psychological and neuroscientific evidence (for reviews see Schultz, 2006; Siegel & Allan, 1996; Ramscar, Dye, & Klein, 2013a), also correctly identify greater variation in the cognitive performance of older adults, and successfully predict that older adults will exhibit greater sensitivity to the fine-grained properties of test items than younger adults. Given that the models run (and can be rerun) on computers, the possibility that any differences in their performance are due to aging hardware can be eliminated; instead, their patterns of performance reflect the information-processing costs that must inevitably be incurred as knowledge is acquired. Once the cost of processing this extra information is controlled for in studies of human performance, findings that are usually taken to suggest declining cognitive capacities can be seen instead to support little more than the unsurprising idea that choosing between or recalling items becomes more difficult as their numbers increase.

2. The problem with "processing speed"

Findings from psychometric tests indicate that the rate at which the mind processes information increases from infancy to young adulthood, and declines steadily thereafter

(Salthouse, 2011). Increasing reaction times are a primary marker for age-related cognitive decline (Deary, Johnson, & Starr, 2010) and are even considered its cause (Salthouse, 1996), yet they are puzzling: Practice improves speed and performance on individual cognitive tasks at all ages (Dew & Giovanello, 2010). Since we continually practice using our cognitive capacities as we age, why does our performance on tests of them decline?

We suggest that the answer to this question lies in the way that psychometric tests neglect learning and its relationship to the statistical patterns that characterize human experience. Learning is a discriminative process that serves to locally reduce the information processing demands associated with specific forms of knowledge and skill (Ramscar, Yarlett, Dye, Denny, & Thorpe, 2010; Rescorla & Wagner, 1972). However, age and experience will inevitably increase the overall range of knowledge and skills any individual possesses, increasing the amount information in (and complexity of) his or her cognitive systems. Processing all this extra information must inevitably have a cost (Shannon, 1948).

2.1. Learning and the long tail of experience

Statistically, the distribution of human experience is highly skewed: Much of our day-to-day life is fairly repetitive, involving a small repertoire of common occurrences, such as reading the newspaper and going to work. At the same time, we encounter a far more diverse range of infrequent or even unique occurrences (as Wittgenstein, 1953, noted, one rarely reads the exact same newspaper twice). When data are distributed in this way, comparisons of means are often meaningless (Baayen, 2001). Consider the problem of recalling birthdays: We are usually reminded of the birthdays of family members on an annual basis, and this usually makes us good at remembering them. However, as we move through life, we learn about other birthdays. Sometimes we hear these dates only once, such as when we attend a party for someone we barely know. As we learn each new birthday, the mean exposure we have had to all the birthdates we know declines, and the task of recalling a particular birthday becomes more complex. Accordingly, it does not necessarily follow that someone who can recall 600 birthdays with 95% accuracy has a worse memory than someone who can recall just six with 99% accuracy.

Psychometric tests do not take account of the statistical skew of human experience, or the way knowledge increases with experience. As a consequence, when these tests are used to compare age groups, they paint a misleading picture of cognitive development. This point can be demonstrated most clearly and effectively in relation to language: It is a central and largely unique aspect of human cognition, and thanks to recent developments in machine information processing, its statistics are more readily and objectively quantified than other aspects of experience. Moreover, almost all psychometric tests involve some form of linguistic information processing: On any test in which subjects have to comprehend verbal instructions and then refer to them in memory in order to perform a task, performance can be influenced by, and may even simply reflect, individual differences in linguistic information loads.

Importantly, linguistic distributions are skewed at every level of description (Baayen, 2001). Consider the relationship between word types (e.g., *dog*) and tokens (how often "*dog*" occurs; Fig. 1). In English, a few words occur very frequently (*the*, *and*), such that half of the tokens in any large natural sample will come from only 100 or so high-frequency types. The relative frequency of these types decreases rapidly (the most-frequent word may be twice as frequent as the second-most), and frequency differences between types decrease as their relative frequency declines. This means that the other half of a large natural sample will be composed of ever-fewer tokens of a very large number of types, with ever-smaller frequency differences between them. Typically, around half of these types occur just once.

This is a very long-tailed distribution: the Corpus of Contemporary American English (COCA; Davies, 2009) contains 425 million entries sampled from a broad range of written sources. Repetitions of the most frequently used 100 words account for 208 million of these entries. The remaining 217 million entries represent 2,800,000 words. Accordingly, although individual low-frequency types are, by definition, rare, their distribution means the chance of encountering a low-frequency token in any sentence is very high (Möbius, 2003).

This distribution ensures both that any English speaker learns only a fraction of the language's total vocabulary, and that individual speakers' vocabularies will grow steadily across the life span. However, the vocabulary tests that are typically used to control for the growth of knowledge in studies of cognitive aging (Salthouse & Mandell, in press) assume vocabulary size is age-invariant in adults (Bowles & Salthouse, 2008; Carroll, 1993; Spearman, 1927), an assumption seemingly confirmed by psychometric vocabulary

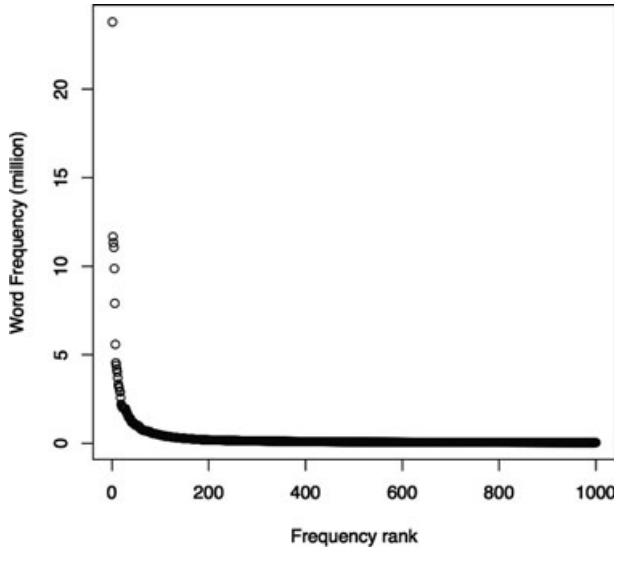

Fig. 1. The frequencies of the 1,000 most common words in the Corpus of Contemporary American English (Davies, 2009) plotted by rank.

measures, which indicate that vocabulary growth in adulthood is marginal, such that increases are only reliably detected in meta-analyses (Verhaeghen, 2003).

Unfortunately, psychometric vocabulary measures are virtually guaranteed to fail to detect vocabulary growth in adults because they attempt to extrapolate vocabulary sizes from sets of test words that are biased toward frequent types (Heim, 1970; Raven, 1965; Wechsler, 1997). However, the distribution of word-types in language ensures both that adult vocabularies overwhelmingly (and increasingly) comprise low-frequency types, and that an individual's knowledge of one randomly sampled low-frequency type is not predictive of his or her knowledge of any other randomly sampled low-frequency type. This makes the reliable estimation of vocabulary sizes from small samples mathematically impossible (Baayen, 2001).

2.2. Simulation Study 1: Why linguistic distributions confound vocabulary estimates

To illustrate these points, we analyzed the statistical properties of a state-of-the-art test designed to measure the vocabularies of advanced learners of English (Lemhoefer & Broersma, 2012). The test samples 40 items (more than most standard vocabulary measures, Bowles & Salthouse, 2008), and its design explicitly seeks to control for the way that the shape of linguistic distributions makes vocabulary measurement a problem (unlike most psychometric vocabulary measures). The upper left panel of Fig. 2 plots the words employed in the test by their rank-frequency in the distribution of English lemmata in the CELEX lexical database (Baayen, Piepenbrock, & Gulikers, 1995; in linguistics a lemma is defined as a "canonical form," such that the frequency of the lemma *walk* comprises the individual frequencies of *walk, walks, walked*, etc.; basing this analysis on lemmata frequencies ensured for more conservative estimates than counting inflected word forms as separate items). As can be clearly seen, all of the types in the Lemhoefer and Broersma test clearly belong to the higher frequency part of the English lexicon: Over half of the lemmas in the CELEX database are lower in frequency than the test items.

To illustrate the way that the distribution of word types affects the measurement of vocabulary growth over time, we constructed a word frequency distribution using the lognormal-poisson model (Baayen, 2001), with parameters estimated from the distribution of English lemmata. We then simulated 20 speakers incrementally sampling from this distribution. (A simplifying assumption made here was that speakers sample the language at the same rate.) The black circles in the upper right panel of Fig. 2 plot increase in vocabulary size with "age" for one simulated learner (because the amount of language individuals are exposed to varies dramatically [Hart & Risley, 1995], for the purpose of these simulations, "age" is defined in terms of the number of lemma tokens a learner has experienced, rather than time). As can be seen, although vocabulary size continually increased in the simulations, the rate at which new lemmata were encountered in the simulations decreased as learners' experience grew. The gray circles then show the vocabulary common to all 20 simulated speakers. This shared vocabulary is typically less than half a speaker's own vocabulary, and further, the rate at which new common lemmata are encountered (i.e., learned) as compared to non-common lemmata decreased

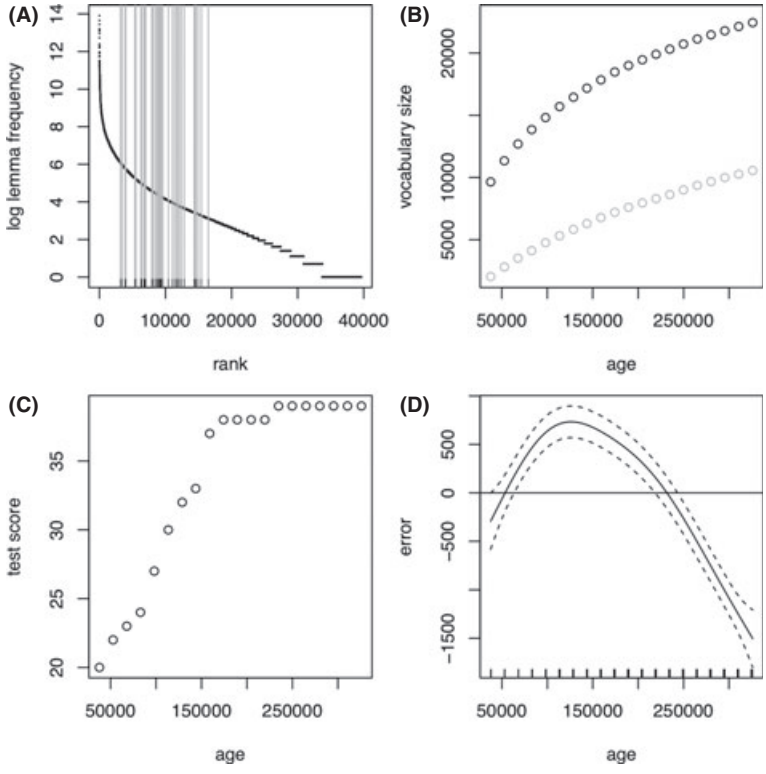

Fig. 2. The non-linear dynamics of vocabulary growth. A: Upper left: the relative frequencies of the Lemhoefer and Broersma (2012) vocabulary test items. Each vertical gray line represents an individual item, and the black curve plots the empirical lemma rank-frequency distribution of English. B: Upper right: the black circles plot an individual vocabulary growth curve, while the gray circles plot the vocabulary that this simulated speaker has in common with the 19 other speakers in the simulation. C: Lower left: the same simulated learner's score on the vocabulary measure over time (because language exposure rates vary dramatically across individuals [Hart & Risley, 1995], age is expressed in these plots as a function of the size of the sample an individual has experienced, rather than time). D: Lower right: error (actual vocabulary – predicted vocabulary) as a function of age.

with age, such that the heterogeneity of the vocabularies in the simulation increased over time.

Consistent with this observation, one of the highly educated reviewers of this article noted that he had never encountered the word *lemmata* prior to reading it here. Given that we can safely assume that the reviewer knows many technical terms that we have not encountered, this may help illustrate why knowing that someone has one very low-frequency word—such as *lemmata*—in his or her vocabulary is of little to no benefit when it comes to predicting whether the person knows another very low-frequency word. It might also, in turn, help illustrate why it is inevitable that any word anyone might reasonably consider including in a "general" vocabulary test will tend to have a fairly high frequency relative to the overall vocabulary of the language, namely because the whole

point of including any particular word in a non-exhaustive measure is to make inferences from it, and very-low frequency words are next to useless in that regard. Because the way words are distributed in language means that most of an adult's vocabulary comprises low-frequency types (with similar distributional properties to the word *lemmata*: highly familiar to some people; rare to unknown to others), this may also in turn help clarify just why the assumption that one can infer an accurate estimate of the size of the vocabulary of an adult native speaker of a language from a small sample of the words that the person knows is mistaken.

This brings us to the bottom left panel of Fig. 2, which plots how scores on the Lemhoefer and Broersma vocabulary test increased as the simulations sampled from the distribution. Initially, test scores rose rapidly, but by middle (and hence older) age, they had asymptoted. The lower right panel then shows that, under the assumption that the test score is a linear predictor of vocabulary size, the Lemhoefer and Broersma test underestimates early vocabulary sizes, then overestimates middle period vocabulary sizes before underestimating the number of lemmata that are learned with still more experience. Accordingly, these test scores are far less sensitive to variance in older vocabularies: A split for $N > 250{,}000$, defining an old speaker group, and $N < 150{,}000$ for the same speakers in their youth, revealed a difference of 1.91 in the variance of the older test scores as compared to 33.22 for the younger scores ($F[159, 103] = 17.44$, $p < 0.0001$); as Fig. 2B shows, while it is empirically the case that the variance in the vocabularies of the older group was greater than that of the younger group, the test scores suggest that the opposite is true.

These results should not be taken to mean that vocabulary measures are useless (e.g., they have a role to play in estimating the progress of language learners; Lemhoefer & Broersma, 2012). What we hope they help make clear is why the insensitivity of vocabulary tests to vocabulary growth in adults is not a sign that vocabulary learning ceases at some point in time. (A moment's reflection on this point might suffice for some readers: As our reviewer's observation shows, it is abundantly clear that even highly educated adults continue to encounter new words on a regular basis throughout the course of their lives.) Given that current studies of aging systematically fail to control for the way vocabulary (and other forms of) knowledge continues to increase throughout adulthood, we next examine the influence that vocabulary growth can be expected to have on cognitive processing.

3. Simulating the effects of vocabulary learning on information processing

Normally developing infants are initially sensitive to all the fine-grained phonetic discriminations made by the world's languages. However, as she learns her native vocabulary, a child's sensitivity to non-native phonetic distinctions diminishes (Werker & Tees, 1984). This is not usually taken to indicate that cognitive decline begins in infancy. Indeed, this loss of sensitivity can be seen as an inevitable result of learning: In discriminative learning, the values of an initially undifferentiated set of cues are shaped

by experience, which drives the discovery of the cue values that best predict a learner's environment (Ramscar et al., 2010; Sutton & Barto, 1998; Rescorla, 1988; Rescorla & Wagner, 1972; Sutton & Barto, 1998). Because this process involves learning to ignore uninformative cues, it explains why decreasing sensitivity to uninformative phonetic information goes hand in hand with increased knowledge of informative phonetic distinctions (Ramscar, Suh, & Dye, 2011).

The learning component of the model we use to simulate the way experience affects reading works in precisely this way. The Naive Discriminative Reader (NDR; Baayen, Milin, Durdevic, Hendrix, & Marelli, 2011; *in sub*) is a two-layer network in which letter unigrams and bigrams (*b, bo, br*, etc.) serve as input cues, and lexemes (the target words that must be discriminated in reading) serve as outcomes. The values of the n-gram cues are initially undifferentiated and are set competitively as the model learns to predict lexemes from the letters it "reads." Every n-gram cue is linked to each lexeme outcome to form a set of subnets, and the cue-weights in these subnets are set by the equilibrium equations of the Rescorla-Wagner learning rule (Danks, 2003). These weights are completely determined by the distributional properties of the model's training corpus, and simulated latencies derived from them capture a very wide variety of empirical effects in reading (Baayen et al., 2011; *in sub*).

3.1. Simulation Study 2: How does vocabulary growth influence lexical decision speeds?

To first examine the effects of adult vocabulary growth on lexical processing, we simulated the effects that experience might be expected to have on lexical decision tasks, in which subjects make a speeded judgment as to whether a letter string is a word or not. There are at least two ways in which experience can be expected to influence this process: First, we might expect that increased experience of any given word will make people better at recognizing it, such they will be quicker to respond to a higher frequency word such as *where* than a lower frequency word like *whelp* (and it is well established that this is the case: Lexical decision responses are slower for lower frequency words than higher frequency, which in turn suggests that lexical decision measures ought not be expected to be experience invariant). Second, we might expect that recognition will not only depend on how often someone has seen any given word but also on the total number of words that he or she knows. Imagine two people, one who knows 20,000 words and another who knows 40,000 words. All other things being equal, we might expect that the first person will be faster at establishing that a newly presented word is in her vocabulary than the second person, simply because she has a smaller space of words to search in through memory (in reality, of course, these two factors will almost certainly interact, since people who have larger vocabularies will almost inevitably also have had more reading experience).

To shed some light on whether the increased vocabulary search that adults encounter as their linguistic experience expands can be expected to have an influence on the speed of their lexical decision responses, we trained two NDR models on data drawn from the Google Ngrams Corpus (Brants & Franz, 2006), a very large, naturalistic data set. (*N*gram

is a term used to describe strings of words and letters in computational linguistics: "you go there" is a word trigram, and xy is a letter bigram.)

We should note at the outset that when it came to training the models, there was and is no clear answer to the question of what the "correct" sample size for each of them might be. The rate at which new words occur in both speech and text varies dramatically according to context and text type (Hayes & Ahrens, 1988), and individual rates of exposure to text are equally variable (Anderson, Wilson, & Fielding, 1988). Accordingly, assumptions about "average readers" must inevitably have an artificial quality. For the purposes of this investigation, we assumed that reading isolated words is largely determined by written frequencies rather than spoken frequencies, and that the reading expertise measured by response speeds in a lexical decision task largely depends on reading experience. To estimate this experience, a conservative reading rate for adults of 85 words/min, 45 min/day, for 100 days/year was adopted. Twenty-one year olds were thus assumed to have 12 years experience reading at this rate, and 70-year olds a further 49 years. Model 1, which simulated reading to age 21 (a typical age for "young adult" subjects in studies), thus "read" 1,500,000 word tokens, and Model 29,000,000 word tokens, simulating reading to age 70 (the typical age for "old adults"). These input parameters were set prior to the simulation results being analyzed.

Furthermore, although the development of large corpora has made simulations of learning from realistic samples of language possible, the automated optical character recognition (OCR) methods used in their construction record a great many misspellings and other spurious items among the unique strings in these samples. Counting these as words is likely to seriously bias estimates of the true distribution of word types in the corpus (Liberman, 2013). To limit the likelihood of spurious items biasing the models' learning, the training sample was drawn from the Google Trigram Corpus. This contains only trigrams with 40 or more occurrences in the Google Corpus, thereby omitting around 50% of its lowest frequency unigram strings (this sample is far more likely to contain OCR errors). To further reduce noise in the models' training, the training sample was further restricted by limiting it to the 14,822,311 trigrams in the Corpus that contain one of the test words used in the English Lexicon Project (Balota & Spieler, 1998).

Sampling from this small subset of the Google Trigram Corpus inevitably meant that the training samples omitted large numbers of legitimate low-frequency strings, such that models encountered far fewer low-frequency words than are likely to occur in a true experiential sample. These very conservative assumptions thus biased the simulations against our hypothesis, since they will tend to result in an earlier asymptote in vocabulary learning than training on the actual underlying distribution.

The input to each network model comprised the letters and letter-bigrams that occurred in each training trigram token. For the trigram "in the box," for instance, the input cues were the letters *i, n*, *t*, *h*, *e*, *b*, *o,* and *x* and the bigrams *#i*, *in*, *n#*, *#t*, *th*, *he*, *e#*, *#b*, *bo*, *ox,* and *x#* (# denotes an orthographic space). Each lexeme in the trigram then served as an outcome, that is, for the trigram "in the box," the outcomes were the lexemes *in*, *the*, and *box*.

3.2. Results and discussion

In keeping with the results of Simulation Study 1, the old model acquired a much larger vocabulary than the young model: The former "learned" 32,536 word types, and the latter 21,307. (These estimates are very conservative: As we noted above, the Trigram Corpus contains only around 50% of the unique string types in the complete Google Corpus. Even with this highly constrained input, vocabulary expansion was far from asymptote after 3 million trigram tokens.)

The empirical reaction times used to evaluate the models were taken from a data set of lexical decision latencies for a set of 2,906 monosyllabic English words ranging from two to eight letters in length constructed by Balota, Cortese, and Pilotti (1999). The latencies were collected for two age groups: younger subjects (mean age: 21.1) and older subjects (mean age: 73.6; the full data set and more information about its acquisition is available at: http://www.artsci.wustl.edu/~dbalota/lexical_decision.html). In a lexical decision task, a string appears on a computer screen, and subjects either recognize it as a word (e.g., *WHELP*) and respond "word," or else fail to recognize it (e.g., *WHERP*) and respond "non-word." To simulate the processing behind these judgments, simulated lexical decision times (SRTs) for each of the words in the Balota et al. test set were estimated as a function of the activations in the NDR models given the ngram cues in the input words (Baayen et al., 2011; a more detailed description of the method is given in the Appendix).

To investigate the frequency by age interaction, we fit a generalized additive model (GAM; Hastie & Tibshirani, 1986) to the data using the R package *mgcv* (Wood, 2006, 2011). The basic structure of a GAM model is:

$$y = X\beta + f_i(x_1, x_2, \ldots) + \ldots + \varepsilon, \tag{1}$$

where y is the response variable, X is a linear predictor, and f_i are smooth functions of the covariates x_k. The parametric part ($X\beta$) of a GAM is identical to that of standard regression models. The non-parametric part $f(x_1, x_2, \ldots)$ consists of a number of smooth functions (f_i) and allows for non-linearities to be modeled more successfully.

The observed and simulated reaction times were then modeled as a smooth on the Google unigram frequencies of the words (Brants & Franz, 2006). The Google unigram frequencies of the items in the test set ranged from 1,337 to 19,401,194,714. To remove rightward skew from the distribution, raw frequencies were log-transformed and any logged frequency values further than 2.5 standard deviations from the mean (0.8% of the data) were removed to prevent overfitting near the edges of the distribution.

Fig. 3 plots the differences in simulated response times (SRTs) between the two models (old–young) over the frequency range of the Balota et al. (1999) items, as well as the same differences for the observed latencies of older and younger adults. The smooths for both were created using the *by* variable of the smooth function in the *mgcv* package (Wood, 2011). To meet the assumption of normally distributed residuals, the GAM models were fitted on inversely transformed reaction times.

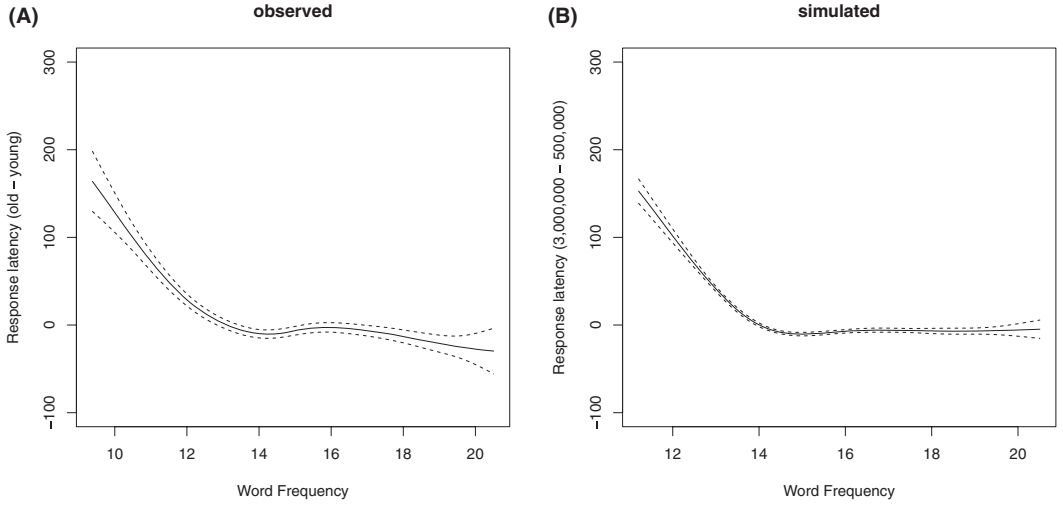

Fig. 3. A: Left panel: fit of a generalized additive model to the response latency *differences* (old–young) across the frequency range between young and old adults for a set of 2,906 English words (Balota et al., 1999). B: Right panel: fit of a generalized additive model to the *difference* in the simulated response latencies (old model–young model) for the same items. The slope on the left side of each plot is caused by differences in the sensitivity of older and younger adults (and the older and younger models) to variance in the lower frequency range of the test items. Whereas the younger adults (and younger model) were insensitive to frequency differences in the lower range, the response times of older adults (and the SRTs of the older model) slowed as the word frequencies at this end of the spectrum declined.

The SRTs and the observed latencies were highly correlated across the frequency range: $r = 0.78$. Furthermore, the models successfully predict an important qualitative difference in the empirical word-frequency effect: While sensitivity to the frequency differences among the test words appeared to asymptote at higher frequencies in both models, frequency sensitivity in the younger model also leveled off at lower frequencies, such that the model only exhibited sensitivity to frequency differences in the mid-spectrum of word frequencies in the test set. In contrast, the SRTs of the older model increase as word frequencies decline across the lower frequency range, such that the *difference* between the SRTs of the older and younger models rises as word frequencies decline (Fig. 3B). This pattern was also found in the empirical data, which, when analyzed, revealed that the older adults were much better attuned to frequency variation in the lower band of the word frequency spectrum than the younger adults (Fig. 3A).

Detailed consideration of the discriminative learning process helps explain these results. In learning, weights on the links between cues and outcomes get adjusted in two ways: Links strengthen when a cue and outcome co-occur, and weaken if cues occur without outcomes. Thus, when "*where*" is encountered, the link between the bigram WH and the lexeme WHERE is strengthened, while, since WH has occurred without WHELP, WH-WHELP is weakened.

In learning, high-frequency words are encountered often, at fairly constant rates (consistently reinforcing WH–WHERE, and weakening WH–WHELP); however, low-frequency

Table 1
The 50 lowest frequency items in the set used to test the models and the older and young adults; BLASH has the lowest frequency of these items, and JEER the highest. As can be seen, many of the letter bigrams in this set of words are comparatively rare in English

1. BLASH	11. CROME	21. TWERP	31. WHELP	41. BLEAT
2. SCHNOOK	12. GIBE	22. THWACK	32. SHUCK	42. CHIVE
3. LETCH	13. LISLE	23. DAUNT	33. MOOCH	43. WHIR
4. ZOUNDS	14. FLAYS	24. RETCH	34. JELL	44. CROON
5. JAPE	15. SPLOTCH	25. GYP	35. GROUCH	45. TAMP
6. SOUSE	16. VELDT	26. YAWL	36. AWN	46. BOSH
7. WHIG	17. SLOE	27. FLUB	37. MANSE	47. RILE
8. FILCH	18. CONK	28. STANCH	38. WRACK	48. BLANCH
9. RHEUM	19. FRAPPE	29. PAUNCH	39. HOOCH	49. LILT
10. PARCH	20. SKULK	30. JOWL	40. FLECK	50. JEER

words only occur sporadically, so WH–WHELP is reinforced and WH–WHERE weakened less often. These imbalances generate "selection pressures" that appear to shape the distribution of cues in the lexicon (see also Zipf, 1949). Thus, the high-frequency items in the test set are shorter ($t[2901] = -10.58$, $p < 0.001$) and have higher bigram frequencies ($t[2901] = 8.98$, $p < 0.001$) than the low-frequency items, which means that low-frequency items contain both more, and rarer, cues (Table 1). Although rare cues have relatively high values in small vocabularies (which is reflected in higher weights in the younger model), they are vulnerable to competition as experience grows, because new vocabulary items will be more likely to share these rare cues. The consequences of this can be seen in the older model: because it has sampled and learned more low-frequency words, the variance in the values of the rarer cues in the older model is greater than in the younger model (see Risse & Kliegl, 2011 for a discussion of how similar factors influence the different reading strategies of older and younger adults).

To ensure older adults' greater sensitivity to low-frequency words was not specific to this particular data set, a second empirical set of data was analyzed in the same way (Yap, Balota, Sibley, & Ratcliff, 2011; Fig. 4). All of the effects reported in the first analysis replicated successfully.

4. Modeling "decline" in a "non-lexical" task

We next examine whether the relationship between information load and response time observed in lexical processing can also be found in tasks termed "non-lexical" in the psychometric literature. It is important to stress here that unless a research protocol can be transferred directly to non-verbal animals (and we know of no psychometric measures that can be), the use of the terms "non-verbal" or "non-lexical" is somewhat misleading: Any procedure that relies on subjects' ability to follow verbal instructions and refer back to them in performing as task must necessarily involve "lexical" processing, and our analysis and results indicate that this processing is likely to be influenced by linguistic

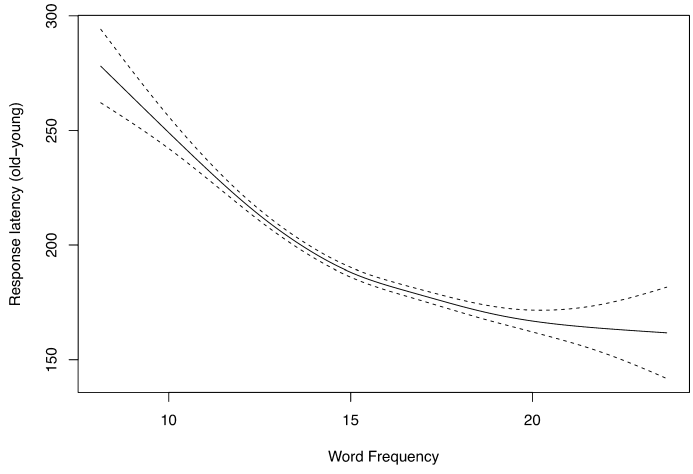

Fig. 4. Average percentile RT differences (old–young) for the naming latencies of 2,820 single-syllable words (Yap et al., 2011) by young (M age: 22.6) and old adults (73.6), plotted against the words' log frequency in the Google 1-gram corpus, and a generalized additive model fit to the RT differences. As with the lexical decision latencies, the rise in the difference slope as frequency descends is driven by the naming latencies of the older adults, which are far more sensitive to differences in the lower frequency range. (The difference in the shape of this slope when compared to one plotted in Fig. 3A likely reflects the different task demands associated with lexical decision judgment tests and naming tests; see e.g., Grainger & Jacobs, 1996.)

experience. This point is obvious in the *letter classification task* (Posner & Mitchell, 1967), the widely used "non-lexical" psychometric test analyzed here, but it is important to note that it seems reasonable to assume that it will apply to the processing of all sensory stimuli in tests where subjects' responses are mediated by instructions that have been communicated to them linguistically.

4.1. Simulation Study 3: How does vocabulary growth affect letter classification speeds?

In the *letter classification task*, subjects are presented with two letters presented in upper or lowercase (A, a, D, d, etc.) and judge whether they represent the same or different alphabetic characters (i.e., *E e* are "same," and *E F* "different"). To simulate behavior in the task, a pair of "old" and "young" NDR models were trained using the methods reported in Simulation 1, and SRTs were used to estimate the empirical latencies for classifying a letter target given the orthographic representation of that letter; that is, *h* was treated as an abbreviated lexeme (to reflect the use of *H* as a symbol for entropy, *R* for a statistical programming environment, *r* for correlation, etc.) and the time to make a classification involving *h* was estimated from the cue strengths for the ngrams *h #h* and *h#*.

4.2. Results and discussion

The observed data used to evaluate the simulation of the letter classification task were obtained from Hale, Lima, and Myerson (1991). The data were obtained using Posner and

Mitchell's (1967) letter classification protocol and were collected for two age groups: Young (M age: 19.6) and Old (69.3). The stimuli comprised five letters, presented in either uppercase or lowercase (*A*, *a*, *D*, *d*, *E*, *e*, *R*, *r*, *H*, *h*). In each trial, two letters were presented simultaneously and subjects were asked to judge if the letters were the same letter of the alphabet. Letter pairs were presented in three conditions: identical (same letter, same case), semi-identical (same letter, different case), or different (different letter).

Older subjects responses are slower than younger subjects (Fig. 5), a finding that replicates straightforwardly in the models once the coupling between letters and their role as abbreviated lexemes is accounted for. (To emphasize how this point plays out in reality, it is worth noting that in the course of their lives, adults will often learn to attach multiple lexemes to a single abbreviated ngram cue; for example, *PFC* is an abbreviation for *prefrontal cortex* in neuroscience, *post-focus compression* in linguistics, and *Private, first class* in military parlance.) The network complexity function employed in calculating the SRTs (see Appendix), which models response latencies as a function of the activation of the lexemes for both letters in a letter pair, predicts longer latencies for older as compared to younger subjects because the larger system of lexical outcomes in the older model makes "accessing" the letter lexemes harder.

Psychometrically, letter classification is often described as an "information-processing" measure, and older adults' longer response times are taken as evidence of declining information processing capacity. Yet information theory—which defines the workings of the information-processing systems that symbolize our age, and which begat the term "information-processing "in the first place—is, at heart, just a set of methods for formalizing the uncertainty in distributions (be they bits of code or vocabulary items; Shannon, 1948). Information is a property of systems, and processing demands are measured in relation to them (MacKay, 2003). In letter classification, the relevant system comprises the task, the subject, and, crucially, what that subject knows. Because psychometric tests neglect this knowledge, they are incapable of measuring information processing in this task (Ramscar & Baayen, 2013). (These points thus echo arguments made in favor of the "rational analysis" of cognitive processes, which consider the properties of the task environment to be essential to understanding human task performance; Anderson, 1990;

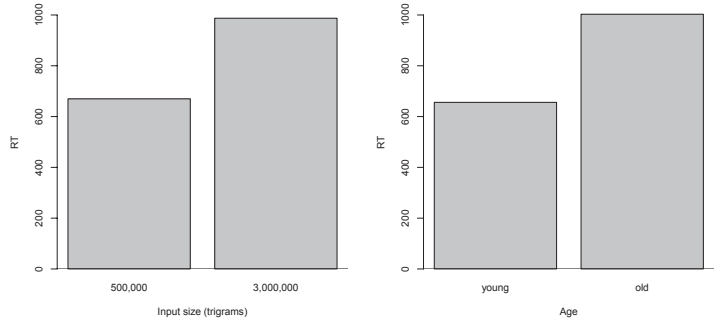

Fig. 5. Letter classification SRTs (left panel, left bar: young model; right bar: old model) and empirical latencies (right panel, left bar: younger subjects; right bar: older subjects).

Anderson & Schooler, 1991; Schooler & Hertwig, 2005; the discriminative learning perspective offered here can be seen as a natural extension of this earlier view in that it emphasizes that functional task environments will themselves vary as a result of learning over the lifetime; see also Ericsson & Kintsch, 1995.)

5. Lexical knowledge and paired-associate learning

All things being equal, one might expect that as an individual's experience grows, his or her knowledge will increase, and that this will in turn raise processing costs in his or her cognitive system (c.f. Shannon, 1948). Consistent with this, our results indicate that slower lexical information processing can simply reflect learning, and that is not necessarily evidence of "decline." Further evidence for this idea comes from the results of comparisons of monolinguals and bilinguals, where an interaction between experience, vocabulary size, and response speed is also observed: The response latencies of young bilinguals in picture-naming tasks resemble older monolinguals more closely than young monolinguals or old bilinguals (Gollan, Montoya, Cera, & Sandoval, 2008). What is notable, however, is that although younger bilinguals exhibit slower response times and increased tip-of-the-tongue rates as compared to younger monolinguals, these differences are not usually thought of as deficits. Rather, the opposite conclusion tends to be reached: Bilingualism is viewed as a cognitive blessing, and bilinguals' lexical processing performance is seen to reflect the natural demands associated with bilinguals' larger vocabularies (Gollan & Acenas, 2004).

However, in the light of our findings so far, the resemblance between the tip-of-the-tongue rates of bilinguals and the elderly raises an intriguing question: Can learning account for age-related memory differences, such as those observed in Paired-Associate Learning[1] (PAL; a psychometric measure of people's ability to learn and recall new information)? In PAL tests, such as the commonly used PAL subtest of Wechsler's Memory Scale (WMS; des Rosiers & Ivison, 1986), subjects learn pairings between word cues (e.g., *baby; jury*) and word responses (*cries; eagle*) and have to supply the appropriate response to each cue at test. Although performance on individual items varies (Fig. 6), on the task overall older adults are slower and less accurate than younger adults, and it has been suggested that this is due to "encoding" (MacKay & Burke, 1990) and "retrieval" deficits in older adults' memory processes (Burke & Light, 1981).

What is not clear, however, is why anyone would expect that PAL performance ought to be age and experience invariant in the first place? First, because performance across the pairs of words used in the test varies, and it seems reasonable to assume that whether a pair seems "easy" or "hard" is itself a function of experience, and second because long-established principles of associative learning predict that well-known words should be harder to learn as Cues (w_1) than less familiar words (Rescorla, 1968), that less familiar words should be easier to learn as responses (w_2) than well-known words (Kamin, 1969), and that w_1–w_2 pairs ought to be hard to learn if w_1 and w_2 occur independently at high rates (Rescorla & Wagner, 1972).

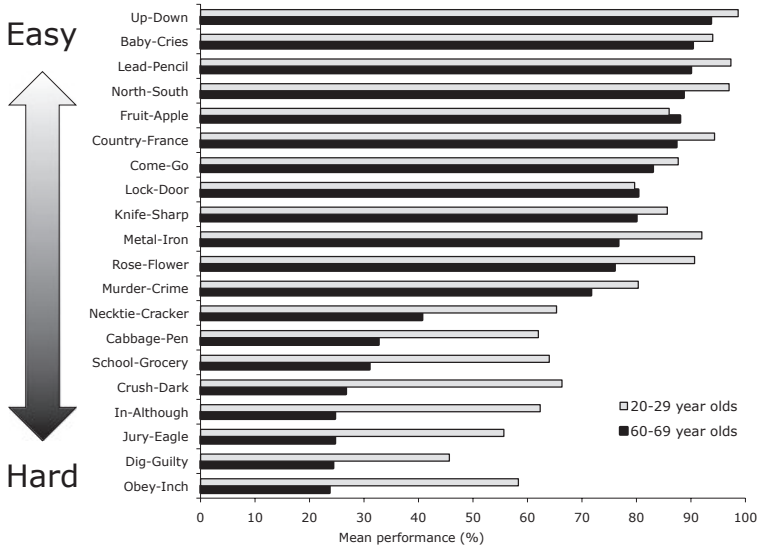

Fig. 6. Mean performance by item for 100 older (age 60–69) and 100 younger (20–29) adults on forms 1 and 2 of the WMS-PAL subtest (desRosiers & Ivison, 1988). As in the lexical decision and naming data, older adults show greater sensitivity to differences in item properties (hard vs. easy) than younger adults.

The WMS classifies the word pairs shown in Fig. 7 as either "easy" or "hard" for testing and scoring purposes: *jury-eagle* is a "hard" PAL pair, and *baby-cries* an "easy" pair. While this makes intuitive sense—and for older adults, at least, this differentiation even turns out to be accurate—it is worth considering what, exactly, might serve to make a given pairing "hard" or "easy" in a learning task. Suppose someone does not know any English: it seems reasonable to assume that in these circumstances, *jury-eagle* and *baby-cries* would be equally easy (or hard) for them to learn. If the person did know some English, then the semantic association between *baby* and *cries* might help the person learn *baby-cries* more easily than *jury-eagle* (Tulving & Pearlstone, 1966). However, as Fig. 7 clearly shows, the greatest point of difference between old and young subjects' performance in PAL learning is on the hard items: Qualitatively, as adult age increases, it would appear that it is the hard items that become harder. While this pattern of performance cannot be straightforwardly explained through appeals to "decline," or by naïve theories of "association," it is easily explained by the principles of learning described above (and which are embodied in the learning model used in the earlier simulations).

A key finding from the study of learning over the past hundred years is that learning is not only sensitive to events that are associated (*the Pavlovian dog learns the bell means that food is going to arrive*) but also, less intuitively, that learning is sensitive to events that dissociate one another: Given one set of cues (*it's a mild day in the hills, and there is not a cloud in the sky*), our cognitive systems actively learn to expect that some events will not occur (o*n hearing a distant rumbling, the idea of thunder is not entertained*; Rescorla, 1968, 1988; Danks, 2003; Ramscar, Dye, & McCauley, 2013b). That is, the

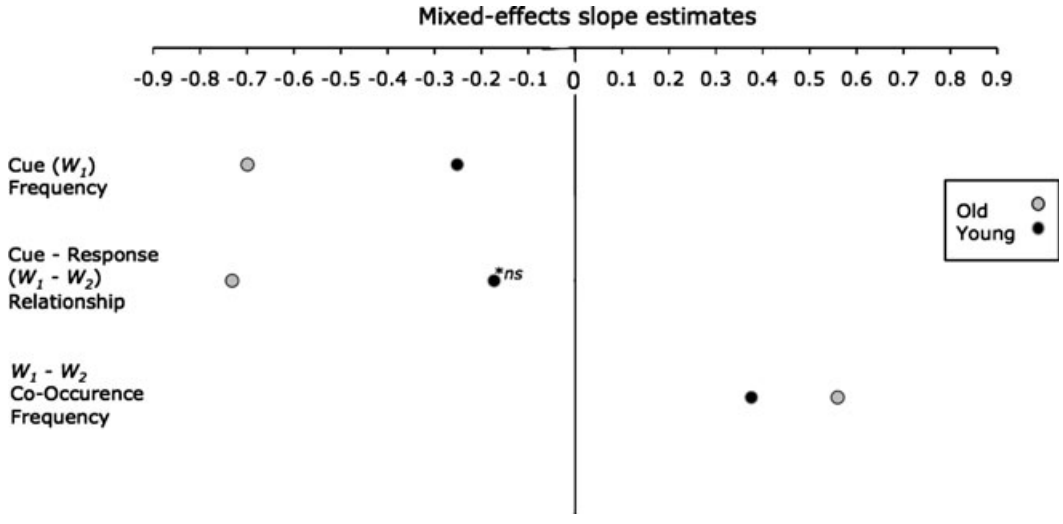

Fig. 7. Mixed-effects slope estimates for the learnability predictors and by-item PAL performance of 60–69 and 20–29 year old adults (desRosiers & Ivison, 1988). All predictor effects and interactions in the model are significant, and all slopes (except *) are significantly different from 0 (t values ≥ 2).

basic principles that guide our best understanding of learning predict that the more often *jury* and *eagle* have been encountered absent one another in language, the harder it will be to learn any subsequent association between them.

5.1. Simulation Study 4: Learnability and paired-associate learning over the life span

To examine whether PAL performance simply reflects factors that predictably influence learning, we analyzed data from a sample of older and younger adults on the WMS-PAL subtest (des Rosiers & Ivison, 1986; both groups comprised identical numbers of subjects and equal numbers of males and females). In a regression analysis, w_1 predictability (log w_1 frequency; t = −4.06, $p < 0.001$), the relationship between w_2 and w_1 predictability (log [w_2 frequency]/log [w_1 frequency]; t = −2.94, $p < 0.01$) and w_1–w_2 co-occurrence rates (log Google frequency; t = 6.77, $p < 0.0001$) accounted for over 75% of the variance in the proportional difference in the item scores (mean old/mean young) of 20–29 and 60–69 year olds (F[3] = 16.4, r = .87, $p < 0.0001$).

We noted above that, all things being equal, the relative learnabilty of w_1–w_2 pairs might be estimated from the co-occurrence and background rates of w_1–w_2. All things are *not,* however, equal: Discrimination learning is a systematic process, and this means that learnabilty itself can depend on experience, which, of course, older adults have more of. As we also noted earlier, because w_2 words will become more independently predictable the more often they are sampled absent w_1, and w_1 words less informative the more they are sampled absent w_2, experience is likely to make learning some w_1–w_2 pairs harder. As experience

grows, PAL performance should increasingly reflect the distributional properties of w_1–w_2 items. Where co-occurrence rates are low, a lifetime of learning that *jury* is uninformative about *eagle* will make learning *jury-eagle* harder, whereas high co-occurrence rates will reduce these effects, making *baby-cries* easier to learn than *jury-eagle*.

A mixed-effects analysis of w_1–w_2 item scores by age confirmed the accuracy of this prediction (Fig. 7). For each predictor, the magnitude of the slope for the older age group is greater than that for the younger age group, indicating that older subjects bring more lexical experience to the task. Consistent with our earlier findings, older adults' PAL performance reflects their greater knowledge of and sensitivity to the distributional properties of w_1–w_2 words, whereas younger adults' less varied performance reflects their more limited knowledge of them. As we noted above, the statistical properties of human experience make comparisons of means invidious: In this case, high average PAL scores ought to be interpreted as measures of ignorance, rather than "intelligence."

6. Names, age, and memory

For many older adults, the problems posed by the task of remembering people's names represent the most disturbing aspect of aging (Cohen & Faulkner, 1986; Lovelace & Twohig, 1990). Given what we have reported so far, this raises a question: To what extent does memory for names really decline, and to what extent are the specific problems that people have with name memory simply a factor of the role that names play in human experience? There are good reasons to believe that names present a unique information-processing problem, and that this problem will be magnified by individual exposure to the distribution of names over time.

First, there is the nature of names as a lexical class: While most nouns are generic—*spoon*, *dog*, *idea*—proper nouns (and especially personal names) are *sui generis*: Ideally, a name uniquely discriminates an individual from her peers. While this could easily be achieved by giving each individual a unique label, this move would *massively* increase linguistic complexity. By now, there would be over a billion English name words, imposing huge demands on lexical processing. Accordingly, languages appear to solve the information problem posed by names by employing name grammars, which form identifiers from smaller sets of hierarchically structured naming tokens (Ramscar, Dye, Gustafson, & Klein, 2013c). Names drawn from a smaller pool of names precede tokens drawn from larger pools of names, and this structure serves to reduce the size of the search problem that speakers and listeners face at any given point in time as names unfold.

As we mentioned earlier, information theory (Shannon, 1948) is a set of formal techniques for quantifying the uncertainty in distributions of discrete events. As such, it provides an objective method for quantifying the search problem posed by systems such as names (Ramscar et al., 2013d), enabling us to measure whether and how name systems may have changed across the lifetime of individual learners, and to estimate the implications of these changes. Moreover, information theoretic methods are discriminative (Shannon, 1956), and thus they provide a means for quantifying the structure of external,

environmental systems, and the dynamics of change in these systems, that naturally complements the insights into internal systems of representation that are provided by discrimination learning models.

Shannon (1948) showed that for coding and measurement purposes, the amount of information provided by events in a discrete system as they unfold in time depends on the way they are distributed: The most efficient distribution is the skewed one seen everywhere in language (Fig. 1), and the least efficient is a "flat" distribution in which all events are equally likely to occur (Shannon, 1948). Intuitively, the reasons behind this can be grasped in relation to names by considering that if 30% of all males are called *John* and only 0.01% are called *Cornelius*, then learning that someone is called *Cornelius* will be more informative than learning he is called *John*. However, although *Corneliuses* will be better discriminated by their names, meaning that hearing *John* leaves a listener more uncertain who, exactly, is being referred to than hearing *Cornelius, Johns* will in all likelihood be easier to remember (guessing *John* will be correct 30% of the time), meaning that knowing that someone is called *Cornelius* will leave a speaker with more uncertainty as to whether she will be able to recall the person's name the next time she needs it than knowing that someone is called *John*.

This example also helps illustrate some of the benefits of, and constraints, on naming systems: The memory advantage enjoyed by a *John* relies on there being lots of other *Johns* in the system, as does the memorability of *Cornelius,* which *Cornelius* also benefits from there being lots of *Johns*. *Cornelius* will be easier to recall if the system has fewer names in total, and would be harder to recall if the *Johns* had a variety of names instead; and, of course, *Cornelius*'s discriminability relies on there not being many *Corneliuses.* This system offers other advantages; for example, everyone will benefit from the fact that the more frequent *John* is shorter and easier to say than *Cornelius*. Simply because the distribution is skewed toward *Johns,* it follows that processing names will, on average, be easier and less time consuming than if the frequencies of *John* and *Cornelius* were reversed (Shannon, 1948; see also Zipf, 1949).

This still, of course, leaves us with the problem of individuating *Johns*, which takes us back to name grammars: Sequentially, if we know someone is called *John,* the search space of subsequent items can be reduced from that of all the surnames we know to just that of the set of surnames that follow *John*. Name grammars thus systematically help balance the competing demands of lexically discriminating between individuals and keeping the speech and memory demands imposed by language processing (Ramscar et al., 2013d). Indeed, the trade-off between the various factors we have described, which act to modulate uncertainty in speakers and listeners as linguistic messages unfold over time, appears to play a large part in shaping the way that linguistic distributions develop and evolve (Ramscar et al., 2010; Ramscar & Baayen, 2013; Jaeger, 2010).

Although name grammars provide speakers and listeners with an external framework that helps considerably in dealing with the information problem posed by names, they cannot eliminate it: In the 2000 U.S. census over 1.15 million different surnames are shared by five or more people; and, consistent with our remarks above about the shape of the distribution of word types, a further 5 million by less than five people (Word,

Coleman, Nunziata, & Kominski, 2008). Even allowing for artifacts of the automated procedures used to read census data, it is clear English has a very large lexicon of surnames. Moreover, although social evolution appears to have produced name grammars that help minimize the information processing challenges posed by names, the widespread legislation of traditional naming practices that accompanied the development of modern bureaucratic states has led to an historically unprecedented growth in size of the pool of first names in English (and many other languages) in the past century, and this has resulted in a concomitant increase in the entropy associated with names (Ramscar et al., 2013d; Scott, 1998). *Entropy*, measured in *bits*, is an information theoretic term used to describe the uncertainty associated with a distribution of events or items, and it provides an abstract, objective measure of the search problem associated with selecting a particular item from a knowledge base containing many items.

However, perhaps because entropy values are a function of the way items or events are distributed—a great many or a very few discrete items could have an entropy of 4 bits depending on their distribution—ideas about thinking about information in terms of *entropy* and *bits* are often difficult to grasp. A common solution to this problem for the purposes of making comparisons in computational linguistics is to convert the entropies of complex distributions of different distributions of items into a measure called *perplexity* (Bahl, Baker, Jelinek, & Mercer, 1977), which expresses abstract *bit* values in terms of a distribution of independent, equally likely outcomes, calculated as 2^H (so that a distribution with an *entropy* $H = 3$ *bits* has a *perplexity* of 8). This allows the uncertainty associated with three bits of information ($H = 3$, which is difficult to grasp), to be expressed in more intuitive terms as the perplexity one would feel if asked to guess "what happens next?" when 8 equally likely outcomes are possible.

Accordingly, while it is of course the case that a great many female names are, and always have been, used in American English, from an information theoretic perspective, the cognitive challenge imposed by recalling a female name in the 1880s can be quantified as being equivalent to that of anticipating a given outcome when a little over 100 equally likely alternative outcomes are possible. As Fig. 8 shows, the perplexity of English first names increased almost exponentially in the years after 1880, such that when recalling a contemporary female name is quantified in terms of perplexity, the comparable challenge can be seen to be equivalent to anticipating a given outcome when over 2,000 equally likely alternative outcomes are possible. This increase is considerable, and given that historically, the entropy of English first names was far lower and far more consistent over time (in the 300 years prior to 1750, 50% of men and 50% of women in England were consistently given one of just three highly frequent male or female names; Smith-Bannister, 1997; Lieberson & Lynn, 2003), it appears that it is also unprecedented.

6.1. Simulation Study 5A: The effect of changing name distributions across time

The very clear change in the entropy of English names that we have described has not been taken account of in any study of name memory that we are aware of, yet it suggests that name processing would have been far easier for 20-year olds in 1960 than it is for

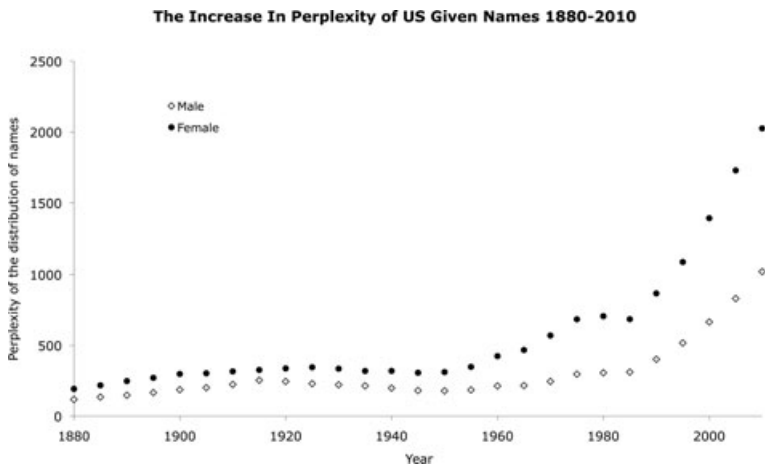

Fig. 8. The *perplexity* of male and female names with a count ≥ 5 in U.S. Social Security applications at 5-year intervals from 1880 to 2010.

20-year olds today, and that the processing load imposed by names has increased dramatically during the lives of today's older adults (Fig. 8). To simulate the effects of these changes on name processing, three NDR models were trained on names sampled at their historical frequencies in 1910, 1960, and 2010 respectively, which were interpolated into an otherwise identical set of naturalistic linguistic training data: 1,500,000 tokens from the Google Unigrams Corpus to simulate the experience of reading to age 20 (because many proper nouns populate the tail of linguistic distributions, sampling from the Trigrams Corpus was likely to result in our considerably underestimating the occurrence rate of less common first names).

The first name frequencies from the historical Social Security application data were converted to appropriate Google unigram frequencies by comparing the ratio of the medians in the Social Security data to those in the Google unigrams. The first names in the Google unigram data were then removed (or else, for tokens such as *June*, which can be either a first name or a month, the Google unigram counts were adjusted to reflect these usage ratios), and replaced by names from the Social Security applications, which were randomly sampled to reflect the distributions appropriate to each period.

After training, SRTs were calculated for each model (see Appendix for more details), both for the set of first names learned by that model, and for the set of names common to the 1910, 1960 and 2010 Social Security application data sets. The first measure allowed for the overall processing imposed by loads names in these periods to be compared, while the second, comparing the SRTs for the names common across the 100-year period represented by the 1910, 1960, and 2010 Social Security applications allowed the effect of changes in the distribution of names on the processing of specific names to be measured. If an SRT is a function of an individual name, then the names common across the 100-year period (1910, 1960, and 2010) should show the same average SRT in each

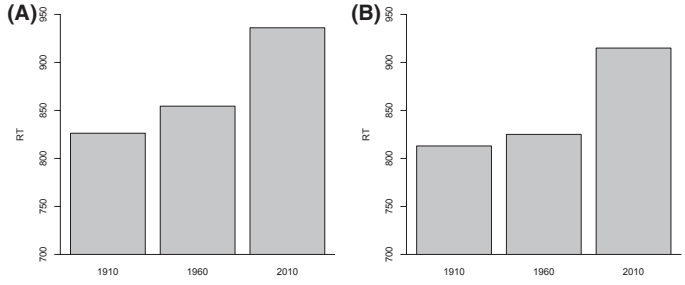

Fig. 9. Left panel (A): average SRTs for the set of first names learned by three 20-year old models trained on 1.5 million unigrams when sampling from the first name distribution in 1910 (left bars), 1960 (center bars), and 2010 (right bars). Right panel (B): SRTs in each model for the set of first names common to the 1910, 1960, and 2010 U.S. Social Security applications: that is, predictions for the same set of first names.

period, even if the overall average SRT were to increase due to an increase in the number of less common names. In contrast, if SRTs are influenced by the greater number of names, then we would expect SRTs for the subset of common names to increase as well.

6.2. Results and discussion

The results of these simulations suggest that the simple task of *recognizing* an American-English first name grew harder in the 20th century: Fig. 9A shows the cost imposed by the total set of names learned by a 20-year old model at each point in time, and Fig. 9B shows the effect this had on the same set of first names (the set of names common to all three periods). The increase in the information processing load imposed by names is especially visible in the latter part of the century: The change in the SRTs from 1960 to 2010 is three times larger than from 1910 to 1960.

Not only was the rise on the number of first names learned dramatic (the 1960 model learned 60% more names than the 1910 model, and the 2010 model 83% more), but the number of non-name words (i.e., words that are not in use also as names) that were learned also declined, by 2.5% in 1960, and 5% in 2010. Given that the models were trained on exactly the same number of first name tokens, this reflects the degree to which the boundary between English first names and non-names has become blurred over time (i.e., *Apple* and *Harmony* and now employed as first names, as well as common nouns), which is likely to further increase the processing problems posed by personal names.

6.3. Simulation Study 5B: The effect of changing name distributions over a lifetime

To simulate the effect that these changes might be expected to have on the life of an individual, we then constructed another 20-year old model, which was trained on 1,500,000 word token sampled from the Google unigrams corpus into which first names sampled from the summed distribution of Social Security applications from 1950 to 1960 (the age at which today's septuagenarians were 20) were interpolated at the rate at which first names occurred in the Google sample using the same procedure as in Simulation Study 5B. We

then compared the model of a 20-year old in 1960 to a second model, which was trained on 9,000,000 tokens sampled from the Google unigrams corpus, into which first names sampled from the distribution from 1950 to 2010 were interpolated. The second thus simulated the effects of 50 extra years' "experience" on the 20-year-old model.

6.4. Results and discussion

The projected impact of name vocabulary growth on lexical processing is shown in Fig. 10A, which plots the impact it can be expected to have on name recognition for someone aged 70 in 2010 as compared to her 20-year-old self 50 years earlier. Fig. 10B shows the projected effect of these processing costs on the same set of names in the same individual (the set of names common to both name vocabularies). In both cases, the model predicts that on average, simply recognizing a name will take today's septuagenarian around half a second longer than when she was 20.

Further insight into the causes of the problems today's older adults experience with name memory comes from an examination of what the models learned in the simulations. Underlining both the degree to which names and other proper nouns comprise a large proportion of the word types in the lexicons of English speakers, and the degree to which this part of the lexicon expands disproportionately with experience, it revealed that whereas the younger model learned 34,480 word types, of which 4,540 were first names, the older model learned 61,839 word types, of which 19,976 were first names. Thus, while the simulation doubled its total vocabulary by age 70, it resulted in a four-fold increase in its first name vocabulary.

The results of these simulations suggest that, given the very real distress name recall causes older adults, the unchecked rise in the information load of personal names we describe should be a cause of social concern; and also, given the objective scale of these changes, that confounding name recall problems with cognitive decline is akin to asking older adults to accept personal responsibility for a social problem.

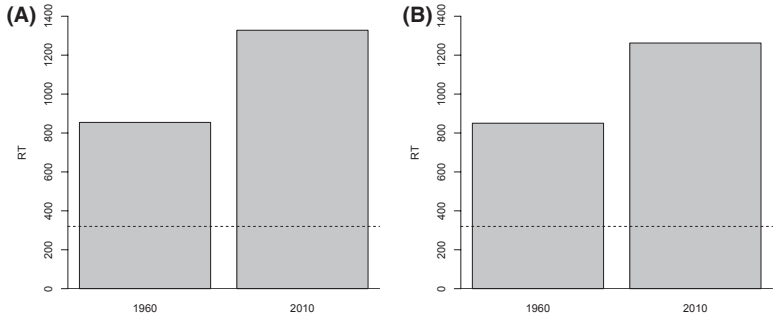

Fig. 10. Left panel (A): average SRTs for the names learned by a 20-year old model (trained on 1.5 million unigrams, including names from the 1950–1960 distributions; left bars), and a 70-year old model (trained on 9 million unigrams, including names from the 1950–2010 distributions; right bars). Right panel (B): SRTs for the names common to 1960 and 2010 U.S. Social Security applications. The area below the dashed line represents a 320 ms response constant (button pressing) added to the SRTs.

7. Can learning explain why some "cognitive abilities" do not decline with age?

Although performance "declines" with age on most psychometric tests, this pattern is not universal. For example, it is generally accepted that "lexical access"—the ability to recall words—declines with age, yet performance on a common lexical access test, the Controlled Oral Word Association Test (COWAT; Spreen & Strauss, 1998) "FAS" subtest, seems to improve with education and experience (Goral, Spiro, Albert, Obler, & Connor, 2007).

In the FAS test, subjects are given 60 s to generate as many words beginning with "F" (then "A," and "S") as they can. Proper nouns and multiple words using the same word-stem (e.g., *friend*, *friends*, *friendly*) are not acceptable. Although studies indicate that older adults can outperform younger adults in this task, often by a large margin (e.g., Czaja, Sharit, Ownby, Roth, & Nair, 2001), on another part of the COWAT test, where subjects are asked to name as many animals as they can in 60 s, studies have shown that performance declines steadily with age (Goral et al., 2007; but see Hargreaves et al., 2011).

Although these contradictory findings have aroused little comment or curiosity in the cognitive aging literature, the idea that the processes that guide the retrieval of animal names might decline while those that guide the retrieval of names beginning with FA or S appear to improve is hard to reconcile with a general decline in retrieval processing. However, as with proper names, it may be possible to account for these patterns by attending to the specific information loads imposed by the tasks: Whereas the animal naming part of the COWAT test simply involves retrieving examples from that distribution of a set of items that are primed by a cue—and one would expect that, to varying degrees, the entropy of the distributions of various classes of nouns and names will increase with experience, thereby increasing the information processing demands associated with tasks involving them—the FAS part of the test is more complex. Subjects are presented with a cue that primes a set of responses but must only report words that are not proper nouns and do not share a "stem" (e.g., *friendship*, *friendly*) with a previously reported word.

Although the entropy of words beginning with F is likely to increase with experience (making F word recall harder), because the set of proper nouns is large, and likely to contain more rare members than other parts of speech, the entropy of proper nouns will probably increase greatest over time than that of other words. Given that, all things being equal, we would expect lower entropy words to be easier to retrieve, increased experience is likely to make the retrieval of valid words in this task easier.

7.1. Simulation Study 6: How and why learning makes a test easier or harder

To examine how these factors might affect FAS performance over time, we annotated each word in a 2.25 billion word corpus of English (UKWaC; Baroni, Bernardini, Ferraresi, & Zanchetta, 2009) with a part-of-speech tag using TreeTagger (Schmid, 1994). To avoid counting strings that represent errors (Liberman, 2013), word types with a

frequency of < 5 in the corpus were discarded. We then took two samples from this corpus in order to get an estimate of the language sample that a "typical" 20-year old and a "typical" 70-year old might be expected to have experienced (based on findings from Hart & Risley, 1995, we assume that our mythical, "typical" English speaker experiences around 10 million words of speech a year).

For each sample, we then calculated the binomial probability that a valid response will be selected at random from the set of all of the types in each lexicon:

$$P(ValidWord) = 1 - P(\mathrm{Pr}operNoun) \tag{2}$$

As Table 2 shows, the binomial probability that an F, A, or S cue will cause a valid response to be selected from the lexicon increases with experience. This is because although proper nouns are the largest set of words in the lexicon, most words we encounter over our lives are not proper nouns, and experience actually increases the probability that adults will retrieve correct responses. Also, as expected, as experience grows, the entropy of proper nouns also increases more than for other word types.

Increased experience will thus increase the influence of both of these factors in favor of selecting valid FAS words. To assess the degree to which these distributional factors actually influence retrievability in the FAS task, we examined their ability to account for empirically observed patterns of performance on the different letters in the task (Hargreaves, Pexman, Zdrazilova, & Sargious, 2012; Tombaugh, Kozak, & Rees, 1999).

To reflect the fact that the majority of valid words produced in the FAS task are common nouns, and that the selection of a valid word will be influenced both by the overall statistical distribution of words and the difficulty of selecting a word of a particular type, the relative production likelihood for words beginning with F, A, and S was calculated as a function of the probability that a valid word would be selected, weighted by the difference in the entropy between proper nouns and common nouns as follows:

Table 2
The average change in the likelihood of selecting a valid response for the letters F, A, and S in the COWAT task after a 200 million word sample and 700 million word sample, and the entropies for the distributions of the different parts of speech in each sample. The increases in bold might be expected to improve performance, while italicized increases might be expected to impair performance

	Probability of Valid Response	Proper Noun Entropy	Common Noun Entropy	Adjective Entropy	Verb Entropy	Other Word Entropy
200 million words	0.43	9.01	8.38	6.43	6.34	2.94
700 million words	0.44	9.37	8.49	6.52	6.36	2.94
Percent change	**2.32**	**3.99**	*1.31*	*1.4*	*0.32*	*0*

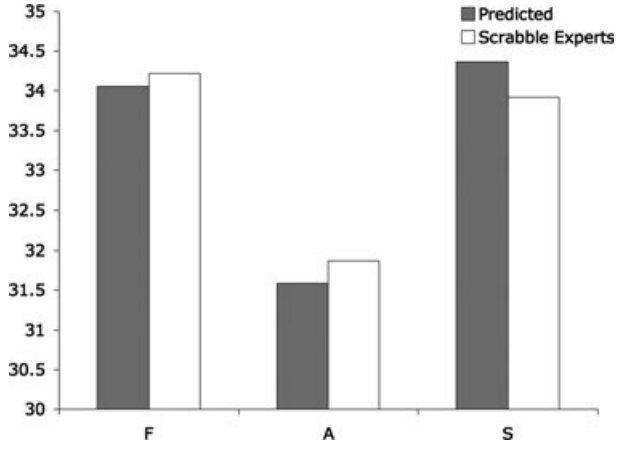

Fig. 11. Proportion of word responses for the letters F, A, and S produced by 23 competitive Scrabble players (M age = 57.2) plotted against the estimates of the difficulty of retrieving valid responses after a 700 million word training sample.

$$P(ValidWord) * (H(ProperNoun) - H(CommonNoun)) \quad (3)$$

For both the 200 million word model and the 700 million word model, these relative production likelihoods correspond well to empirically observed patterns of performance (F, S > A). Furthermore, the fit between the performance of a group of Scrabble experts (M age = 57.2 years; Scrabble imposes a set of task demands that are highly similar to the FAS test in that players must choose valid words given letter cues, but proper nouns are classed as invalid words) and our estimate of the relative difficulty of F, A, and S (based on the 700 million word model) was encouraging (Fig. 11). It is worth adding that in the Hargreaves et al. (2011) study, the Scrabble experts produced 76% more words than undergraduates (M age = 19.4), but only 40% more than the age-matched controls, and that on average, both the undergraduates and the age-matched controls found generating A words most difficult, with S slightly easier than F, as predicted by the model.

7.2. Discussion

As we noted earlier, the patterns of performance observed on the different parts of the COWAT test are incompatible with the idea that "retrieval processes" decline with age in any straightforward way. They are, however, compatible with the growth of information in the lexicon in the animal naming part of the test, and the interaction between competing lexical factors in the FAS part of the test.

It appears that information processing in the animal naming test may simply get harder if the information load in people's cognitive systems grows, when more animal names (i.e., *cat, dog ... shar pei; bonobo; meerkat,* etc.) are learned over time (Goral et al., 2007; but see Hargreaves et al., 2011), whereas in the FAS task, the growing information

load associated with vocabulary acquisition will be offset by the decline in the extent to which the invalid items associated with the retrieval cues serve to interfere with the recall of valid items, such that the task becomes easier over time.

8. A "meta" meta-analysis of FAS performance

Although the results of many studies show that older adults often outperform younger adults in the FAS task (see e.g., Hargreaves et al., 2012), a meta-analysis of 134 studies employing the COWAT FAS task (Barry, Bates, & Labouvie, 2008; also the CFL task in which letters C, F, and L are used) found that performance does decline with age. To reconcile this with our understanding of FAS task demands, we re-examined the meta-data (presented in full in Barry et al., 2008), discovering a relationship between sample sizes in studies and performance (Fig. 12) that had not been previously analyzed.

A median split of the metadata showed that in studies with smaller samples (M = 21.8), performance (M = 42 vs. M = 38.4) was better than in studies with larger samples (M = 303.4; t(132) = 3.65, $p < 0.001$). A linear mixed effects regression analysis underlined the significance of this relationship. When added to the original predictors (Table 3), Sample Size, along with Education and Age accounts for over 50% of the variance in the metadata (stepwise backwards model comparison indicated that none of the other predictors was significant).

To visualize and examine for non-linearities in these interactions, the metadata were modeled in a generalized mixed additive model (GAMM; Fig. 13). As can be seen, in small, "artisanal" samples (10–20 subjects), performance changes very little over time: It increases up to age 30, and then plateaus. Furthermore, while Education is strongly predictive of performance in the metadata, it is negative correlated with age (r = −.34). It is

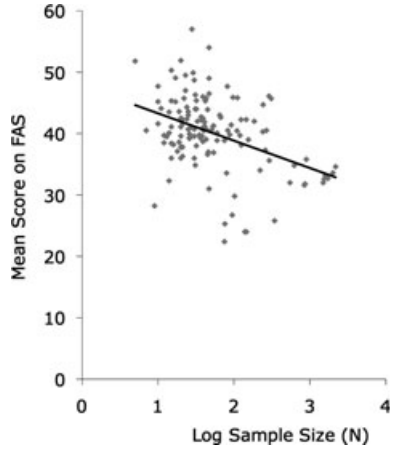

Fig. 12. Mean performance on the FAS/CFL tests in 134 studies (Barry et al., 2008) plotted by sample size.

Table 3
LMER analysis (*Mean Score ~ [Sample Size + Education] * Age*) of mean performance by study on the Barry et al. metadata. Sample Size was added to the original predictors (Year of study, % Male, Form [FAS vs. CFL], Age, Exclusion Criteria, and Years on Education), and then insignificant predictors were removed by stepwise backwards model comparison (13 observations were omitted due to missing values in the metadata)

	Estimate	SE	*t* value	*p*
(Intercept)	3.443	0.443	7.778	0.000
N	0.334	0.116	2.872	0.005
Education	−1.955	0.381	−5.136	0.000
Age	−1.304	0.266	−4.896	0.000
N:Age	−0.212	0.067	−3.153	0.001
Education:Age	1.368	0.224	6.109	0.000

Notes. Multiple $R^2 = 0.54$, Adjusted $R^2 = 0.52$; $F(5,115) = 26.7$, $p < 0.0001$.

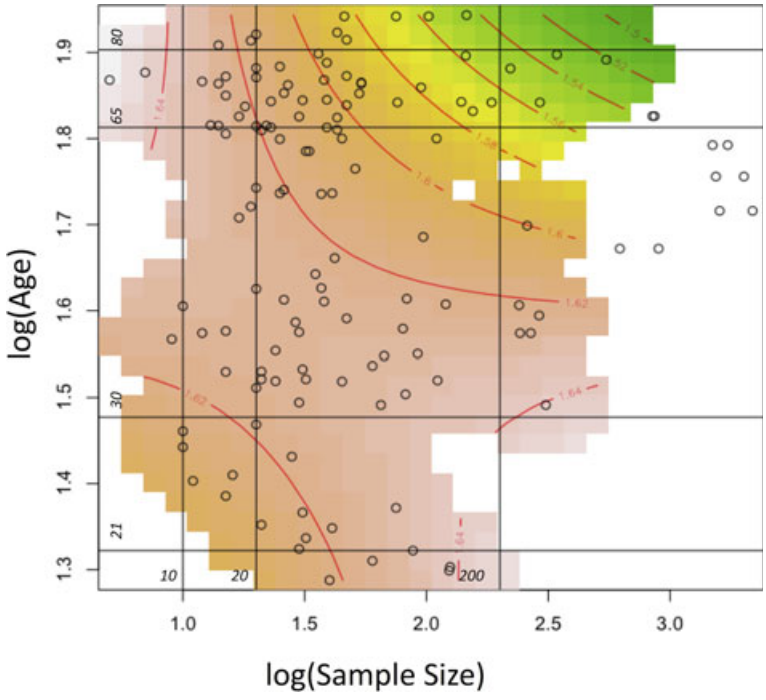

Fig. 13. GAMM tensor smooth of the Barry et al. (2006) metadata. The model (*Mean.Score ~ te(SampleSize, Age) + te(Age, Education)*) takes account of the same factors as the lmer model above. Performance is plotted by log Age (*y*-axis) and log Sample size (*x*-axis) as a heat contour, with better performance in red (points represent the mean performance in each study sample). The superimposed lines are for ages 21, 30, 65, and 80, and samples of 10, 20, and 200. As can be seen, performance only declines with age as sample sizes grow.

thus worth noting that in a post hoc comparison of studies with small samples ($N \leq 40$), the performance of subjects aged 35 and under and 65 and over was indistinguishable (Fig. 14; $t(48) = 0.35$, $p > .7$).

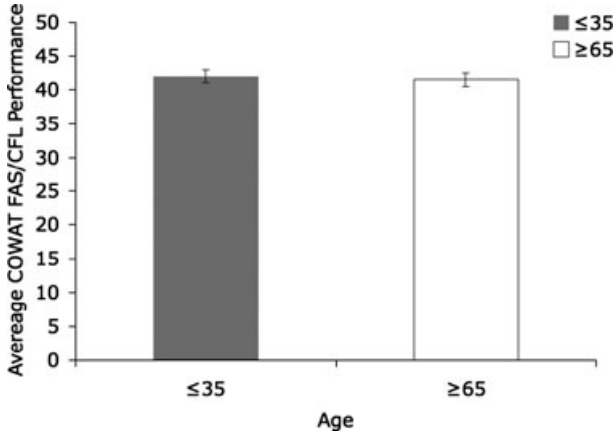

Fig. 14. Average FAS/CFL scores in smaller samples ($N \leq 40$) in the Barry et al. (2008) metadata for subjects aged 35 and under (21 studies, 69% of total: range 19.35 – 34.09, SD = 5.3; M Sample Size = 23) and 65 and over (29 studies, 42% of total: range 65–85.3, SD = 5.4; M Sample Size = 22.1). Error bars are SEM.

8.1. Discussion

The results of this reanalysis of the Barry et al. (2008) metadata are consistent with the pattern of performance predicted by the information structure of the task and the environment, as well as the results of many experimental studies in the literature in which sample size has been controlled for; for example, the FAS performance (M = 48.8) of 23 age-matched controls (M age = 57.4) for the Scrabble experts in Hargreaves et al. (2001) was vastly superior to that (M = 38.6) of 23 college-aged controls (M age = 19.4).

This last result also suggests an answer to a critical question posed by this analysis: *How can sample size influence performance on a test as simple as naming words beginning with F?* The answer may be straightforward: Psychology is a social science. Psychological tests cannot be assumed to have the objective status of, say, litmus tests.[2] Whereas many chemical tests are (or can be considered to be) impervious to social context, it is unlikely that any psychological test is. As studies grow larger, for example, they may be more likely to be run by less experienced or less confident experimenters, which may influence performance. Furthermore, even the behavior of skilled experimenters may change as sample sizes grow: Exclusion criteria might be applied differently to subject 497 in a sample of 800 as compared to subject 17 of 20 (if so, some part of the "decline" seen as test samples grow may simply be an artifact of undiagnosed dementia).

Importantly, because FAS performance is age invariant, we were able to identify sample size as a potential confound. Given that there is no reason to believe the 134 studies analyzed here are not representative of the literature, it is likely that other measures have been similarly affected, adding further distortion to our understanding of aging.

9. Learning and cognitive maturation

The results reported here indicate that older and younger adults' performance in psychometric testing are the product of the same cognitive mechanisms processing different quantities of information: Older adults' performance reflects increased knowledge, not cognitive decline. In discussing this finding, a question continually arises: *"Learning appears to predict linear patterns of change, but cognitive decline really kicks in at around 60 or 70: how do you explain this?"*

In answering this, we first note that as people age, it has been found that they encode less contextual information in memory (Naveh-Benjamin & Old, 2008). Although this is usually taken to indicate that the processes that "bind" contextual information in memory decline with age, learning theory predicts that experience will increasingly make people insensitive to a great deal of background context, simply because ignoring uninformative cues is an integral part of learning (Kruschke, 1996, 2001, 2005; Ramscar et al., 2013a; Rescorla, 1968).

Learning is also sensitive to the environment, and its predictions change with it: If a common environmental change like retirement was to systematically reduce the variety of contexts people encounter in their lives, learning theory predicts that the amount of contextual information they learn will drop further, as the background rates of cues in remaining contexts rise (Kruschke, 1996; Ramscar et al., 2013a). It follows from this that if people were to increasingly spend time in environments where any cues have high background rates already (family homes), any effects arising from their cumulative experience of learning to ignore task irrelevant contextual (background) cues will be exacerbated. In other words, because discriminative learning by its very nature reduces sensitivity to everyday context (Kruschke, 1996; Ramscar et al., 2013a; Rescorla & Wagner, 1972), retirement is likely to make memories harder to individuate and more confusable, absent any "cognitive declines," simply because retirement is likely to decrease contextual variety at exactly the time when the organization of older adults' memories needs it most.

Well-established principles of learning thus explain both the changes that are often perceptible in older adults' cognitive performance around retirement age and the fact that these changes are not detected in testing. In contrast, claims about "cognitive decline" are descriptive, and our findings strongly suggest that these descriptions are erroneous and serve only to perpetuate myths (see also Baltes & Schaie, 1974).

Unfortunately, this does not mean that the diseases that can undermine cognition in old age are similarly mythical, which raises a second question often put to us: *What about all the neurobiological evidence for cognitive decline?* Our answer is that except in the case of neurological diseases where there is evidence of pathology, there is no neurobiological evidence for any declines in the processing capacities of healthy older adults. Although many claims to the contrary have been made (see Morrison & Baxter, 2012, for a review), it is important to note that absent a model of what is being processed, and how, neurobiological studies can reveal only that the structure and/or biology

of neural processing changes; interpreting this as evidence of decline (or increased efficiency) requires a model of the relationship between neural activity and cognitive *function.*

Even if our own explanations of cognitive processing still leave questions unanswered, the contrast with theories that interpret neurobiological changes as decline is stark: At present, most functional accounts of cognitive processing in relation to the brain amount to little more than metaphors (e.g., Park & Reuter-Lorenz, 2009), and even where "computational" models are offered, they overwhelmingly take high-level programming languages as their inspiration, ignoring the information-theoretic constraints that govern the compiled programs that are actually run on physical information processing systems (Ramscar & Baayen, 2013).

Accordingly, while we acknowledge that the models we have presented here are abstract and may not offer much by way of insight into the specific ways by which the brain gives rise to the mind, we believe the discriminative approach employed here has a much to offer: first, because it allows us to formally generate falsifiable predictions at a functional/behavioral level (which are often surprisingly accurate; Ramscar et al., 2010, 2011, 2012, 2013a, 2013b, 2013c, 2013d) and second, because it requires "folk" ideas about cognition to be recast in terms that are more compatible with the workings of physical information processing systems (Ramscar & Baayen, 2013).

Finally, although we have focused on one well-understood learning mechanism in this article, we should note that human learning is not the product of just this one process: It is abundantly clear, for example, that learning is influenced by social as well as environmental factors, and that self-perception can exert a strong influence on what is actually learned from the environment (Dweck, 1999). Because of this, the ideas about "cognitive decline" we have critiqued here are likely to be exerting a strong, negative influence on the lives of many millions of older adults. We hope this can change. Formal models of learning and information processing offer practical as well as scientific insights, and a better, more widespread understanding of these ideas can help people manage their memories more effectively in the future. At the outset, we noted that population aging is seen as a problem because of the fear that older adults will be a burden on society; what is more likely is that the myth of cognitive decline is leading to an absurd waste of human potential and human capital. It thus seems likely that an informed understanding of the cognitive costs and benefits of aging will benefit all society, not just its older members.

Acknowledgments

This research was funded in part by an Alexander von Humboldt research award to Harald Baayen. We are grateful to Denis Arnold, Melody Dye, Wayne Gray, Thomas Hills, Mike Jones, Rheinhold Kliegl, Mark Liberman, Bradley Love, Tim McNamara, Robert Port, Rich Schiffrin, Fabian Tomaschek, and Chris Turnbury, who, along with two anonymous reviewers, provided us with many helpful comments on these ideas.

Notes

1. We thank Rich Shiffrin for this suggestion.
2. While the findings we report are related to concerns raised in the psychology literature about the danger of interpreting significant statistical findings based on small samples, they differ in that there is no "right" number of responses in the FAS task (subjects are simply asked to write down as many words as they can), nor are there any experimental manipulations (subjects' ages are a given). Thus, many of the questions about the appropriate number of subjects to sample in order to draw valid conclusions in experiments, etc., that have been raised recently (e.g., Ferguson & Heene, 2012; Francis, 2012) are not directly relevant in this instance. To the extent that these findings are indicative of a broader trend, it may be that in studies of human behavior, while small samples come with a greater risk of false positive findings, larger samples may in turn run the risk of returning false negative findings; and that in the case of comparisons *between* small and large samples, the chances of either a false negative result or, as it seems here, a false positive result can both increase, depending on the specifics of the mismatched samples.
3. In Simulation Study 2, the accuracy of the models was improved when weights w_1 and w_2 in (8) were estimated separately for the old and young groups, suggesting that the two groups find a slightly different balance between cue-outcome learning (V/a_i) and network size management $(f(V))$. The first weight was estimated at 0.024 for the older subjects, and at 0.029 for the younger subjects. This is consistent with our suggestion that network management costs are greater for older adults than for younger adults.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, *2*, 396–408.

Anderson, R. C., Wilson, P. T., & Fielding, L. G. (1988). Growth in reading and how children spend their time outside of school. *Reading Research Quarterly*, *23*, 285–303.

Attwell, D., & Laughlin, S. B. (2001). An energy budget for signaling in the grey matter of the brain. *Journal of Cerebral Blood Flow Metabolism*, *21*, 1133–1145.

Baayen, R. H. (2001). *Word Frequency Distributions*. Amsterdam: Kluwer.

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

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database (CD-ROM)*. Philadelphia, PA: Linguistic Data Consortium.

Bahl, L., Baker, J., Jelinek, E., & Mercer, R. (1977). "Perplexity—a measure of the difficulty of speech recognition tasks." In Program, 94th Meeting of the Acoustical Society of America 62:S63, Suppl. no. 1.

Balota, D. A., Cortese, M. J., & Pilotti, M. (1999). Item-level analyses of lexical decision performance: Results from a mega-study. Abstracts of the 40th Annual Meeting of the Psychonomics Society (P. 44). Los Angels, CA: Psychonomic Society.

Balota, D. A., & Spieler, D. H. (1998). The utility of item level analyses in model evaluation. *Psychological Science*, *9*, 238–240.

Baltes, P. B., & Schaie, K. W. (1974). The myth of the twilight years. *Psychology Today*, *7*(10), 35–40.

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation 43*(3): 209–226.

Barry, D., Bates, M. E., & Labouvie, E. (2008). FAS and CFL forms of verbal fluency differ in difficulty: A meta-analytic study. *Applied Neuropsychology*, *15*(2), 97–106.

Brants, T., & Franz, A. (2006). *Web 1t 5-gram version 1*. Philadelphia: Linguistic Data Consortium.

Bowles, R. P., & Salthouse, T. (2008). Vocabulary test format and differential relations to age. *Psychology and Aging*, *23*(2), 366–376.

Burke, D. M., & Light, L. L. (1981). Memory and aging: The role of retrieval processes. *Psychological Bulletin*, *90*, 513–546.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge, UK: Cambridge University Press.

Charles, S. T., & Carstensen, L. L. (2010). Social and emotional aging. *Annual Review of Psychology*, *61*, 383–409.

Cohen, G., & Faulkner, D. (1986). Memory for proper names: Age differences in retrieval. *British Journal of Developmental Psychology*, *4*, 187–197.

Czaja, S. J., Sharit, J., Ownby, R., Roth, D. L., & Nair, S. (2001). Examining age differences in performance of a complex information search and retrieval task. *Psychology and Aging*, *16*(4), 564–579.

Danks, D. (2003). Equilibria of the rescorla-wagner model. *Journal of Mathematical Psychology*, *47*(2), 109–121.

Davies, M. (2009). The 385 + million word corpus of contemporary american english (1990-present). *International Journal of Corpus Linguistics*, *14*, 159–190.

Deary, I. J., Johnson, W., & Starr, J. M. (2010). Are processing speed tasks biomarkers of cognitive aging? *Psychology and Aging*, *25*, 219–228.

Deary, I. J., Corley, J., Gow, A J., Harris, S. E., Houlihan, L. M., Marioni, R. E., Penke, L., Rafnsson, S. B. & Starr, J. M. (2009). Age-associated cognitive decline. *British Medical Bulletin*, *92*, 135–152.

Dew, I. T. Z., & Giovanello, K. S. (2010). The status of rapid response learning in aging. *Psychology and Aging*, *25*(4), 898–910.

Dweck, C. S. (1999). *Self-theories: Their role in motivation, personality and development*. Hove, UK: Psychology Press.

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, *102*(2), 211–245.

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, *7*(6), 555–561.

Francis, G. (2012). The psychology of replication and replication in psychology. *Perspectives on Psychological Science*, *7*(6), 585–594.

Gollan, T. H., & Acenas, L. A. (2004). What is a TOT? Cognate and translation effects on tip-of-the-tongue states in Spanish-English and Tagalog-English bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 246–269.

Gollan, T. H., Montoya, R. I., Cera, C. M., & Sandoval, T. C. (2008). More use almost always means a smaller frequency effect: Aging, bilingualism, and the weaker links hypothesis. *Journal of Memory and Language*, *58*, 787–814.

Goral, M., Spiro, A., Albert, M., Obler, L., & Connor, L. (2007). Change in lexical retrieval skills in adulthood. *Mental Lexicon*, *2*(2), 215–240.

Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review*, *103*(3), 518–565.

Hale, S., Lima, S. D., & Myerson, J. (1991). General cognitive slowing in the nonlexical domain. *Psychology and Aging*, *6*(4), 512–521.

Hargreaves, I., Pexman, P., Zdrazilova, L., & Sargious, P. (2012). How a hobby can shape cognition: Visual word recognition in competitive Scrabble players. *Memory & Cognition*, *40*(1), 1–7.

Hart, B., & Risley, T. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: Brookes.

Hargreaves, I. S., Pexman, P. M., Zdrazilova, L., & Sargious, P. (2012). How a hobby can shape cognition: Visual word recognition in competitive Scrabble players. *Memory & Cognition*, *40*, 1–7.

Hastie, T., & Tibshirani, R. (1986). Generalized additive models (with discussion). *Statistical Science*, *1*(3), 297–318.

Hayes, D. P., & Ahrens, M. (1988). Vocabulary simplification for children: A special case of 'motherese'? *Journal of Child Language*, *15*, 395–410.

Heim, A. W. (1970). *AH 4 group test of general intelligence*. London: NFER-Nelson.

Hentchel, H. G. E., & Barlow, H. B. (1991). Fininding minimum entropy codes with Hopfield Networks. *Network*, *2*, 135–148.

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

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In: B. Campbell & R. Church (Eds.), *Punishment and aversive behaviour*. New York: Crofts.

Kruschke, J. K. (1996). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *22*, 3–26.

Kruschke, J. K. (2001). Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology*, *45*(6), 812–863.

Kruschke, J. K. (2005). Learning involves attention. In. G. Houghton (Ed.), *Connectionist models in cognitive psychology* (pp. 113–140). Hove, East Sussex, UK: Psychology Press.

Lemhoefer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, *44*, 325–343.

Lennie, P. (2003). The cost of cortical computation. *Current Biology*, *13*, 493–497.

Liberman, M. (2013). String frequency distributions [Web log article]. Retrieved from Language Log, http://languagelog.ldc.upenn.edu/nll/?p=4456 February 3, 2013.

Lieberson, S., & Lynn, F. B. (2003). Popularity as a taste. *Onoma, Journal of the International Council of Onomastic Sciences*, *38*, 235–276.

Lovelace, E. A., & Twohig, P. T. (1990). Healthy older adults' perceptions of their memory functioning and use of mnemonics. *Bulletin of the Psychonomic Society*, *28*, 115–118.

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms*. Cambridge, UK: Cambridge University Press.

MacKay, D. G., & Burke, D. M. (1990). Cognition and aging: A theory of new learning and the use of old connections. In T. Hess (Ed.), *Aging and cognition* (pp. 213–263). Amsterdam: North-Holland.

Möbius, B. (2003). Rare events and closed domains: Two delicate concepts in speech synthesis. *International Journal of Speech Technology*, *6*, 57–71.

Morrison, J., & Baxter, M. (2012). The ageing cortical synapse: Hallmarks and implications for cognitive decline. *Nature Reviews Neuroscience*, *13*, 240–250.

Naveh-Benjamin, M., & Old, S. R. (2008). Aging and memory. In. J. H. Byrne, H. Eichenbaum, R. Menzel, H. L. Roediger, & D. Sweatt (Eds.), *Learning and memory: A comprehensive reference* (pp. 787–808). Oxford, UK: Elsevier.

Park, D. C., & Reuter-Lorenz, P. A. (2009). The adaptive brain: aging and neurocognitive scaffolding. *Annual Review of Psychology*, *60*, 173–196.

Posner, M. J., & Mitchell, R. F. (1967). Chronometric analysis of classification. *Psychological Review*, *74*, 392–409.

Ramscar, M. J. A., & Baayen, R. H. (2013). Production, comprehension, and synthesis: A communicative perspective on language. *Frontiers in Psychology*, *4*, 233. doi:10.3389/fpsyg.2013.00233.

Ramscar, M., Dye, M., Gustafson, J. W., & Klein, J. (2013c). Dual routes to cognitive flexibility: Learning and response conflict resolution in the dimensional change card sort task. *Child Development*, *84*(4), 1308–1323.

Ramscar, M., Dye, M., Popick, H. M., & 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., Dye, M., & Klein, J. (2013a). Children value informativity over logic in word learning. *Psychological Science*, *24*(6), 1017–1023.

Ramscar, M., Dye, M., & McCauley, S. (2013b). Error and expectation in language learning: The curious absence of 'mouses' in adult speech. *Language*, *89*(4).

Ramscar, M., Smith, A. H., Dye, M., Futrell, R., Hendrix, P., Baayen, R. H., & Starr, R. (2013d) The 'universal' structure of name grammars and the impact of social engineering on the evolution of natural information systems. In. M. Knauff, M. Pauen, N. Sebanz & I. Wachsmuth (Eds.), *Proceedings of the 35th Meeting of the Cognitive Science Society*, Berlin, Germany.

Ramscar, M., Suh, E., & Dye, M. (2011) How pitch category learning comes at a cost to absolute frequency representations. In. L. Carlson, C. Hölscher & T. Shipley (Eds.), *Proceedings of the 33rd Meeting of the Cognitive Science Society*, Boston, MA.

Ramscar, M. J. A., Yarlett, D. G., Dye, M. W., Denny, K., & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, *34*(6), 909–957.

Raven, J. C. (1965). *Guide to using the Mill Hill vocabulary test with progressive matrices*. London: HK Lewis.

Rescorla, R. A. (1968). Probability of shock in the presence and absence of CS in fear conditioning. *Jouranl of Comparative & Physiological Psychology*, *66*, 1–5.

Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American Psychologist*, *43*(3), 151–160.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory*. New York: Crofts.

Risse, S., & Kliegl, R. (2011). Adult age differences in the perceptual span during reading. *Psychology and Aging*, *26*(2), 451.

des Rosiers, G., & Ivison, D. (1988). Paired-associate learning: Normative data for differences between high and low associate word pairs. *Journal of Clinical Experimental Neuropsychology*, *8*, 637–642.

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review.*, *103*, 403–428.

Salthouse, T. A. (2009). When does age-related cognitive decline begin? *Neurobiology of Aging*, *30*, 507–514.

Salthouse, T. A. (2011). Consequences of age-related cognitive declines. *Annual Review of Psychology*, *63*, 5.1–5.26

Salthouse, T., & Mandell, A. R. (2013) Do age-related increases in tip-of-the-tongue experiences signify episodic memory impairments? *Psychological Science*, published online 8 October 2013, doi: 10.1177/0956797613495881

Schmid, H. (1994). *Probabilistic part-of-speech tagging using decision trees*. Manchester, UK: Proceedings of International Conference on New Methods in Language Processing.

Schooler, L. J., & Hertwig, R. (2005). How forgetting aids heuristic inference. *Psychological Review*, *112*(3), 610–628.

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

Scott, J. C. (1998). *Seeing Like The State*. New Haven, CT: Yale University Press.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Systems Technical Journal*, *27*(3), 379–423.

Shannon, C. E. (1956). The bandwagon. *IRE Transactions on Information Theory*, *2*(1), 3.
Siegel, S., & Allan, L. G. (1996). The widespread influence of the Rescorla–Wagner model. *Psychonomic Bulletin and Review*, *3*, 314–321.
Singh-Manoux, A., et al. (2012). Timing of onset of cognitive decline: Results from Whitehall II prospective cohort study. *British Medical Journal.*, *344*, d7622.
Smith-Bannister, S. (1997). *Names and naming patterns in England, 1538–1700*. Oxford, UK: Oxford University Press.
Spearman, C. (1927). *The abilities of man*. New York: Macmillan.
Spreen, O., & Strauss, E. (1998). *A compendium of neuropsychological tests: Administration, norms, and commentary*. Oxford, UK: Oxford University Press.
Sutton, R., & Barto, A. (1998). *Reinforcement learning*. Cambridge, MA: MIT Press.
Tombaugh, T. N., Kozak, J., & Rees, L. (1999). Normative data stratified by age and education for two measures of verbal fluency: FAS and animal naming. *Archives of Clinical Neuropsychology*, *14*(2), 167–177.
Tulving, E., & Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior*, *5*(4), 381–391.
Verhaeghen, P. (2003). Aging and vocabulary score: A meta-analysis. *Psychology and Aging*, *18*, 332–339.
Watkins, K., et al. (2005). *International cooperation at a crossroads-aid*. UN Development Programme: Trade and Security in an Unequal World.
Wechsler, D. (1997). *Wechsler Adult Intelligence Scale–Third Edition (WAIS-III)*. San Antonio, TX: The Psychological Corporation.
Werker, J., & Tees, R. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*, 49–63.
Wittgenstein, L. (1953). *Philosophical investigations*. London: Blackwell.
Wood, S. N. (2006). *Generalized Additive Models*. New York: Chapman.
Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, *73*:3–36.
Word, D. L., Coleman, C. D., Nunziata, R., & Kominski, R. (2008). Demographic aspects of surnames from census 2000. Unpublished manuscript, Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.192.3093&rep=rep1&type=pdf (May, 2012).
Yap, M., Balota, D., Sibley, D., & Ratcliff, R. (2011). Individual differences in visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *38*, 53–79.
Zipf, G. K. (1949). *Human behavior and the principle of least-effort*. Philadelphia, PA: Addison-Wesley.

Appendix

Simulated lexical decision times (SRTs) for each of the words in the Balota et al. test set were estimated as follows: First, the activation *A* of a lexeme given its orthographic form was calculated by summing the weights from the input cues to the lexeme. For the word *car*, for instance, this involved summing the cue strengths of the letters *c*, *a,* and *r* and the bigrams *#c*, *ca*, *ar,* and *r#* to the lexeme *car*.

To model the effect of experience on lexical processing in the two empirical subject populations, we make three further assumptions. First, we assume that the total input activation to the set of cues is the same across populations of different ages, such that in reading, a cue-set will receive the same input activation from the perceptual system irrespective of age. Given that older subjects have larger vocabularies than younger subjects,

this will have consequences for the amount of activation that an outcome receives. Similarly, our second assumption is that, analogous to Kirchhoff's current law (the principle of conservation of electric charge), a principle of conservation of activation holds, such that the amount of activation arriving at cue j (I_j) is equal to the amount of activation spreading from that cue to its associated lexemes (Attwell & Laughlin, 2001; Lennie, 2003 provide evidence that these assumptions are consistent with the observed character of neurological processing).

Accordingly, if we let M denote the set of meaning outcomes, and O_{ji} denote the activation spreading from cue j to meaning i, then

$$I_j = \sum_{i \in M} O_{ji} \tag{4}$$

If V denotes the cardinality of M, we have under uniformity that

$$O_{ji} = I_j \ /V = 1/V \tag{5}$$

for any cue j given unit activation for cue j. As a consequence, the activation a of a lexeme updates to

$$a_i = \sum_{j \in C} \frac{1}{V} w_{ji} \tag{6}$$

$$= \frac{1}{V} \sum_{j \in A} w_{ji},$$

where C is the set of active cues present and w is a weight parameter that determines the final activation of the target lexeme relative to the size of the rest of the system.

Finally, in line with other models of cognitive processing, we assume that the brain is a physical information-processing device, and that increased vocabulary size will alter the effective channel capacity of the networks over which vocabulary-related activation spreads. This is because Shannon entropy, which sums across the transformed probabilities of outputs, is a function of both V and the shape of its distribution, and the *source coding theorem* (Shannon, 1948) proves that (1) *H(V)* defines the lower bound of the code rate (the average number of bits per symbol) in a noiseless system (i.e., the Shannon entropy of the source represents the minimum code length required to discriminate all of the symbols in the source, below which information will be lost) and (2) that a coding scheme is most efficient if, on average, messages are equal to *H(V)* bits in length, which in turn means that the size of the most efficient scheme necessarily increases as V increases (see Hentchel & Barlow, 1991 for a discussion of the application of this to neural coding schemes).

Accordingly, we consider that two factors will contribute to the change in the channel capacity of a network as vocabulary size grows: First, any increase in the number of outputs from an ensemble of neurons will result in an increase in the lower bound of the code rate of the source; Second, the increase in the lower bound of the code rate will

produce a concomitant decrease in the amount of redundancy that is encoded in a signal of any given length: that is, in a signal comprising b bits, if the average number of bits needed to encode a message m increases, then the average number of redundant bits r will decrease:

$$r = (b - m) \tag{7}$$

Increasing the vocabulary encoded in a network must therefore increase both the length of the code processed by the network *and* channel noise across it (Shannon, 1948). However, weights in the individual subnets of the NDR model (which estimate the relative discriminability of items, rather than their probabilities) are not sensitive to information gain in the overall system. Accordingly, a non-decreasing function *f(V)* was entered into the specification of the model to allow for the effects of increased processing and signaling complexity across a growing system of subnets to be factored into our reaction time estimates.

Reaction times for each age group were thus estimated as:

$$RT_i = w_1 V/a_i + w_2 f(V) + c \tag{8}$$

where *f(V)* serves as a network complexity function that reflects the information load on NDR as a function of the number of subnets in the system, and c is a response execution constant.[3]