Dilution as a model of long-term forgetting

Mark Lansdale
School of Psychology, Leicester University
Thom Baguley
Division of Psychology, Nottingham Trent University

Address for correspondence: Mark Lansdale, School of Psychology, Leicester University, Leicester, LE1 9HN.

Tel.: +44 (0)116 229 7121
Fax.: +44 (0)116 229 7196
Email: ml195@le.ac.uk
Dilution as a model of forgetting

Abstract

We develop a model of long term forgetting based upon three ideas: i) Memory for a stimulus can be described by a population of accessible traces; ii) The probability of retrieval after a delay is predicted by the proportion of traces in this population that will be defined as correct if sampled; and iii) This population is diluted over time by null traces that, if accessed, block retrieval. We model this dilution as a linear function of time and the outcome of accessing memories by their temporal organisation. We apply our model to five published experiments studying forgetting in cued recall, four recognition experiments, and one using savings methods. The model specifies the different form of the retention function in each case well, and provides a principled explanation for some puzzling characteristics of forgetting without recourse to mechanisms such as decay or consolidation.
A recent article by Wixted (2004) is one of many on the nature of the forgetting function. In the last 20 years this has included articles by: Loftus (1985), Bogartz (1990), Loftus and Bamber (1990) Wixted and Ebbeson (1991), Laming (1992), Rubin and Wenzel (1996), Rubin, Hinton and Wenzel (1999), and Brown, Neath and Chater (2007); among many others. The mathematical specification of the forgetting function endures as a central theoretical issue because it has proven difficult to resolve convincingly and because it is seen as an important constraint upon how fundamental processes of memory are modelled. In his article, Wixted specifically argues that the significance of Jost’s Law\(^1\) to this debate has been neglected. This law states that of two memory traces of equal strength at a given time, the younger trace will decay more rapidly than will the older. This is illustrated in Figure 1 in which two stimuli, A and B, are learned at different times and to different degrees, with Stimulus A being the earlier and receiving more learning trials. Typically, the retention function for Stimulus B learned at the later time declines faster and eventually intersects that of A; with Jost’s Law, as stated above, applying clearly at that point of intersection. For the purposes of this article, the significance of Wixted’s argument is that the available evidence constrains the possible theoretical explanations of Jost’s Law to models that assume that the proportional likelihood of forgetting decreases with the age of the trace. Rather than being a footnote in the history of memory research, or trivial in the sense that it follows merely from the curvilinear nature of the forgetting function, Wixted demonstrates that Jost’s Law reflects something fundamental about memory.

Insert Figure 1 about here
Dilution as a model of forgetting

We use this starting point to present a simple idea about memory that provides a new explanation of Jost’s law and, in so doing, makes broader claims about the characteristics of the retention function. To this end, we make three points. First, in our analysis of the retention function, we support Wixted’s assertion that the proportional rate of forgetting declines with the age of the trace; although there are some important differences in how we arrive at this conclusion and what theoretical possibilities it might entail. Second, in the spirit of widening the theoretical debate, we offer a new mechanism for this; distinct from the list of possibilities considered by Wixted. We argue that forgetting can be readily modelled as a simple and constant linear function of time, but that the dynamics of memory are such that it manifests itself as a proportional decrease in the probability of access with increasing delay. Third, in making that argument, we comment upon the relationship between the empirical analysis of forgetting functions and the theoretical context in which analysis takes place, to reassert the point, sometimes overlooked, that both elements are essential in formulating a coherent account of the forgetting process (see also Laming, 1992; Wickens, 1999).

The structure of the paper is as follows. We begin by describing the theoretical context within which our assumptions about forgetting are grounded. We are interested in how multiple trace models of memory in which populations of traces – including some that, if accessed, generate errors or block further attempts at retrieval – are used to describe the outcome of learning. To this we add the simple assumption that forgetting is a process in which this population is diluted by non-functional (null) traces over time. We then apply this idea to nine well known studies to illustrate both its ability to model data gathered from a range of methodologies, and its capacity to
Dilution as a model of forgetting

model characteristics unexplained by previous models. We begin first with the
analysis of five cued recall experiments. These allow a simple implementation of our
Population Dilution (PD) model and demonstrate the generality and potential of the
basic idea. We then present an extension of the same modelling approach to four
recognition experiments where the requirement to model different recognition
methodologies, and characteristics of false alarm rates, requires us to develop a more
general and complex implementation of population dilution as a model of forgetting.
We do this by fusing the stochastic processes of the earlier versions of the PD model
with elements of Signal Detection Theory. Finally, we present a simulation of
forgetting as measured by savings measures to further demonstrate the generality of
the approach. This is followed by a general discussion of the scope of the model and
some issues arising from it.

Theoretical context: recall and forgetting in a mixed population model
An extended analogy serves as an advanced organiser for the theory we specify
below. This analogy, and the following theoretical exposition, is aimed at specifying
the recall process. The extension of these ideas to recognition and savings is given in
a later section where the demand of modelling recognition processes requires further
specification of the model.

Suppose, earlier in their career, a pirate has buried several separate chests of
treasure on a long beach. Returning some years later, the pirate, forgetting exactly
where the chests are buried, carries out a geophysical survey of a sufficiently long
stretch of beach to be sure of including the chests. The survey identifies many
possible signals of variable strength that could be treasure chests or they could be odd
Dilution as a model of forgetting

bits of buried flotsam and jetsam. Unfortunately, there will only be one chance to dig a hole and get away before attracting unwanted attention. What is the chance of digging up a chest as opposed to an old bucket?

In this analogy of a cued recall experiment, the signals represent a multiplicity of mnemonic traces and cued recall is a one-shot process of selecting one target from a population of potential targets; including unwanted items. In detail, the theoretical mechanism for recall and forgetting we propose develops previously published models of memory (Lansdale & How, 1996; Lansdale, 2005) and shares the following characteristics whose correspondence to our analogy is obvious: (a) Multiple representations in memory; (b) null, irrelevant or erroneous representations; (c) no resampling of memory upon retrieval failure; and (d) variation in the strength of traces. Before describing our model of forgetting, it is useful to elaborate the background to this theoretical framework:

(a) Multiple representations. Most learning and memory experiments generate instances in which stimuli are experienced more than once. This can arise from multiple presentations, as in experiments concerned with learning and overlearning, or from overt or covert rehearsal. It is not a new idea to suppose that each instance generates a distinct memory record. Although such models are relatively uncommon in the literature, there are a number of strong demonstrations of their explanatory power (e.g., Bower, 1967; Hintzman, 1986; Logan, 1988; Laming, 1992; Brown et al., 2007). They also have the potential for relating the frequency of events in the outside world to the structure of memory; an issue of evolutionary significance (Anderson & Milson, 1989; Anderson & Schooler, 1991).
(b) Null, irrelevant, or erroneous representations. In our analogy, memory traces are represented by buried artefacts that include junk (that is, things that would be a mistake to retrieve) as well as valid targets. Evidence that this can occur in memory experiments is clear. Many experiments induce errors from the participant that can be demonstrated to endure in memory and to compete with recall of correct information (e.g., Loftus, Miller & Burns, 1978; Lansdale & Laming, 1995). Prior associations between experimental stimuli and extra-experimental information can also be seen as representations that can potentially be accessed and recorded as errors later in an experiment. Errors or inaccurate representations (by which we mean any representation that fails to deliver an operationally-defined ‘correct’ response) can also arise as the by-product of normal encoding processes (Lansdale, 2005). There are many ways in which memory experiments make it possible for participants to encode and retrieve erroneous information;

(c) Null traces and the failure to resample memory. Extending the previous two propositions, we argue that retrieval is a sampling from a population of traces made available in given experimental circumstances. That overall population size $P$ is comprised of a subset $C$ that will elicit a response defined as ‘correct’ and a subset $E$ that will likewise generate an error. Lansdale and How (1996) further demonstrated the need to add a third subset of $W$ traces (such that $P = C + E + W$) whose overt effect is to block recall. We refer to these as ‘null’ traces. They could be void recalls (from which nothing useful is recalled) or they could represent errors that the participant knows are incorrect and therefore censors. For the construction of a model
Dilution as a model of forgetting

of recall, their importance is that accessing from subset $E$ and $W$ means the participant will not make a correct response.

On this basis, the probability of correct recall, $p(R)$, at time $t$ is directly predicted by the ratio of correct traces in the overall memory population $C/P$ (begging for the moment the question of how $C$ and $P$ are actually quantified). Mixed population models of this kind (so named to emphasise both the multiplicity of representation and the varying composition of correct, erroneous and null traces) have been used to predict not only the levels of correct recall for different degrees of learning and overlearning in an experiment studying the learning of sequences, but also the nature of forgetting a week later as a function of learning history (Lansdale & How, 1996). This included accurate frequency predictions for occasions where recall occurred at levels significantly below what would otherwise be expected by chance. This is explained by the recall of errors when they outnumber correct traces in the memory population.

Implicit in this view is a further assumption that resampling of memory is inhibited when an error or a null trace is accessed. That is, cued recall is effectively a one-shot process; or at least, repeated attempts at recall cannot be seen as independent of the outcome of earlier attempts. This idea has recurred in the memory literature (e.g., Barnes & Underwood, 1959; Tulving & Hastie, 1972; Reason & Lucas, 1984; Baguley, Lansdale, Larkin & Lines, 2006) although the basis of this may not be well understood (Nickerson, 1984);
Dilution as a model of forgetting

\((d)\) Variation in trace strength. An overwhelming body of evidence points to the view that memory has a quality of strength (alternatively termed ‘familiarity’ and ‘evidence’ amongst other terms) that is expressed in the confidence with which participants make responses in memory experiments and the probability of their being correct. In many memory experiments participants are explicitly required to make decisions on the basis of this. For example, in a two-alternative forced-choice recognition experiment, the task is to determine which of two stimuli has been seen before, and this is often seen as a decision in which the strengths or familiarity of the target and distractor items are compared. The theoretical substrate of strength may be complex in the sense that it could be represented in memory in many different ways, but its elements need not be mysterious. There is no difficulty in supposing, for example, that one way in which traces differ is in the extent to which they represent contextual information that reliably associates the current testing environment to prior experimental situations. In short, we can reasonably assume that populations of traces in a multiple representation model will vary in strength, and the question for this paper is whether, and how, this impacts upon models of forgetting.

If memory strength is treated as a continuous dimension, it is clear that memory processes based upon it are statistically complex and require appropriate models. Recognition memory in particular, where the covariation of probabilities of correct recognition and false alarm rates is compelling, is well-suited to modelling with Signal Detection Theory (e.g., see Wickens, 2002). Naturally, in applying our Population Dilution model to recognition memory experiments below, we are compelled to address similar issues. How SDT is applied in detail is currently a topic of controversy (e.g., see Wixted 2007; and see further discussion below). In this
Dilution as a model of forgetting

In this paper, we have the additional complication of modelling mixed populations of traces emanating from the same stimulus rather than a single entity, and this too has its impact upon how considerations of strength are assimilated into the PD model. As it happens, for reasons detailed below, these considerations are not directly relevant to the modelling of cued recall data, with which we deal first. For clarity, we return to issues of strength in the later section in which we apply our model to recognition experiments.

Forgetting as dilution

We now present an elaboration of this previously published mixed population model to apply it to the forgetting function and Jost’s Law. Suppose the constant $W$ is, in fact, a simple function of time; the simplest being $W = kt$. In other words, the total number of traces $(C + E + W)$ in a given population expands as a linear function of time. Recall as a function of time therefore takes the general form

$$p(R) = \frac{C}{C + E + kt}$$  \hspace{1cm} [1]

where $C$, $E$, and $k$ are constants characteristic of the trace population and $t$ is the time elapsed between the recall test and the stimulus presentation. We refer to this as a population dilution model of forgetting because the effect of this expansion process is to dilute the proportion of correct memories in the population as time passes. In practice, we demonstrate later in this article that this function is an approximation (albeit a very useful one) of a more complex process. However, for clarity of exposition we examine cued recall with this simpler function before presenting the more complex model for recognition.
Dilution as a model of forgetting

Equation [1] predicts Jost’s Law of forgetting because although two populations representing memories of different ages might share the same probability of recall at a given time, they will have rates of change depending upon the size of their trace populations. This arises because of the proportional impact of the \( k \) traces introduced per unit time. Younger, less overlearned, items, such as Stimulus B in Figure 1, will have smaller trace populations resulting from fewer learning opportunities. Therefore the effect of the constant influx of null traces, compared to the larger populations emanating from older, more overlearned stimuli such as A, is proportionately greater. This logic is examined in more detail in Appendix A, where we also elaborate (for later reference) upon the non-linear nature of the retention function predicted by the Population Dilution model when plotted on logarithmic scales, as is commonly done (e.g., Anderson, 2000). This non-linearity represents a characteristic by which the data can differentiate between this model and commonly assumed rivals such as the power law.

Although we use Equation [1] as a mathematical description of forgetting to be tested against known data, it is useful to specify its theoretical elements in more detail here. We present Equation [1] as equivalent to arguing that recall is an outcome of three psychological processes: (a) access to a segment of the temporal record via its temporal organisation; (b) a filtering of that segment to include only those traces reaching threshold levels of association with the cues for recall; and (c) a random selection from the remaining population. We illustrate these processes (again restricting ourselves at this stage to discussing cued recall) in Figure 2.

Insert Figure 2 about here
(a) Access by temporal organisation. Following Glenberg and Swanson (1986), Brown et al. (2007) and others, we assume that cued recall is mediated by a process of temporal discrimination. In the PD model, a segment of the memory record is partitioned to include only those memories temporally indistinguishable from the stimulus to define the population of traces from which recall is attempted. The width (i.e., duration in the memory record) of this segment is modelled as a constant proportion of the overall elapsed time $T$ between presentation and testing according to a simple application of Weber’s Law. We assume the segment is centred upon the stimulus, and therefore all traces encoded $\Theta T$ seconds before and after the stimulus are temporally indistinguishable from it. This defines the width of this segment as $2\Theta T$ seconds; where $\Theta$ is the Weber Constant (e.g., see Baddeley, 1986; Laming 1986). For the purposes of this paper, we approximate $\Theta$ at 0.05 (see Laming, 1986, Table 5.1 for estimates bracketing this value). Thus, just as successful discrimination of relative recency increases with absolute delay between two items (Yntema & Trask, 1963; Bjork & Whitten, 1974), so in the PD model the number of traces that cannot be distinguished from the stimulus by temporal discrimination increases linearly with increasing delay between presentation and test. Because this temporal partition includes the time of the experiment itself, we assume a constant number of correct and error traces (represented by $C$ and $E$ in Equation 1) have been encoded as the result of the experimental procedure. Note however, that it is possible that additional representations of $C$ and $E$ may have occurred extra-experimentally in such a way as to fall within the segment of the memory record accessed. This would also occur if participants rehearsed what they recalled after the experiment. The effect of
Dilution as a model of forgetting

this will be a linear increase in value of $C$ and $E$ with increasing delay, giving a more complex retention function of the form

$$p(R) = \frac{(C + k_c t)}{(C + E + t(k + k_e + k_c))} \quad [2];$$

where the parameters $k_e$ and $k_c$ reflect the density of $C$ and $E$ in the memory record outside of the stimulus presentation. For simplicity of initial presentation and testing of the model (and because it does not prove necessary at this stage) we disregard this possibility here and reconsider this issue in the General Discussion, below.

(b) Filtering by strength of association to the recall cue. The second process in the PD model is that the trace population so defined by temporal cues is further filtered to admit only those traces reaching a threshold. This threshold reflects the degree of association with the cues for recall in a given memory test and is a necessary assumption to model the specificity of recall cues. The effect of this process is to exclude, as far as possible, traces irrelevant to the current cues for recall and thereby to maximise the likelihood of recall. The inexactness of such processes allows for some exclusion of valid traces and some inclusion of null traces. Thus the constants $C$ and $E$ in Equation [1] admit of the possibility that not all eligible traces survive a filtering process and the proportion $kt$ reflects those null or irrelevant traces that do.

We do not elaborate on this filtering process here because in our modelling of cued recall experiments the constants $C$, $E$ and $k$ are estimated as free parameters of the data; which is why we are also able to disregard Equation [2] as a model of the retention function at this stage.
Dilution as a model of forgetting

(c) Random selection of traces. Once a population of traces has been defined by temporal discrimination and filtering, retrieval is modelled as a random choice from that mixed population with the likelihood of correct recall specified in Equation [1].

The scope of this article. Before describing the application of the PD model to published datasets below, it is important to specify the range of memory phenomena we aim to address and those we do not. We are principally concerned with demonstrating the ability of the PD model to model the retention function in cued recall and recognition experiments in terms of the proportion of traces in a mixed population that, if accessed, will support correct recall. We therefore pay particular attention to studies varying the number of learning events on the retention function. These manipulations specifically influence the composition of the trace population as expressed in Equation [1] and therefore allow stronger tests of our model.

Within this broad aim, there are three specific aspects of forgetting we do not model here, for the following reasons. First, we apply the PD model only to data recording forgetting in long term cued recall and recognition studies. Laming (1992) demonstrated that principles similar to those above (in which recall is sampled from a multiplicity of traces) can be applied to short term memory experiments. He did not consider longer term experiments, nor the consequences of reproducing erroneous or null information from memory – although he did consider the interfering effect of other trials in short term experiments that can act in a similar way. In contrast to that work, the model we propose applies in circumstances in which participants do not (or are unable to) use devices such as maintenance rehearsal or temporal recency that make some items more available for retrieval than others. We therefore do not model
Dilution as a model of forgetting

short term forgetting nor consider, beyond some simple observations, the relationship between long and short term forgetting processes (although we return to question the validity of excluding short term memory in a later section). We suspect that short term forgetting is, in some respects at least, more complex than in the long term, because the very nature of short term tasks, the opportunities for rehearsal, and the significant role of temporal cues allows a greater diversity of competing information and ways participants can approach the tasks. This view – based essentially on the idea that memory can be used in different ways in the short and long term – is sustainable even if a multiple storage model of long and short term memory is not assumed (e.g., Laming 1999; Nee, Berman, Sledge Moore & Jonides, 2008). Therefore, we justify studying long term forgetting only in order to establish some simple models against which forgetting in the short term can be usefully contrasted to inform further theoretical development in the future.

Second, for similar reasons, although we model cued recall and recognition experiments, we do not consider free recall. Recent research (e.g., Ward, Woodward, Stevens & Stinson, 2003; Laming, 2006), points to the relationship between the number of rehearsals and recall that may well be modelled by multiple trace models. As noted above, Laming (1992; 2006) also considers the loss of accessibility with increasing lag in terms closely related to the approach we take here. However, much of the debate in the modelling of free recall centres upon the sequence of recalls and the balance of the effects of stimulus repetition (mediated by rehearsals) and recency. These issues span the boundary of short and long term recall that we have chosen to avoid for the reasons given above. Furthermore, given that much rehearsal in free recall is covert, one of the principal explanatory mechanisms of the PD model – the
Dilution as a model of forgetting

numerosity of stimuli and its link to the composition of memory populations – is itself unconstrained and largely a matter of conjecture. Since our purpose in this article is to examine the explanatory power of a very simple idea with as few additional assumptions as possible, the complexity of free recall therefore places it beyond the scope of the current article.

Third, the present paper is concerned with showing how the statistics of populations changing as a function of time can be used to model the forgetting function in a number of experiments; this being a substantial enterprise in its own right. An extension of these ideas to model response latencies and the dynamics of retrieval is an important and logical next step. However, in order to focus upon the central theoretical ideas we are exploring, and given the scale of the analyses we present, we do not develop these issues beyond some general remarks in the final section.

Finally, we comment that ‘dilution’ in the PD model is defined only in relative terms. In Equation [1] above, $k$ is the product of two constants: the Weber Constant, and the proportion of the total number of null traces in the segment of the memory record that are accessible by the current cues for recall. If the density of null traces accessible to the current cues for recall after filtering is $D$ per unit time (representing a fraction of all traces in the temporal record), then $k = 2\theta D$. In the analyses below, Equation [1] is often expressed relative to $C$. That is, $p(R) = 1/(1 + E/C + kt/C)$. In addition, since the PD model estimates recall as the proportion of $C$ traces in a population, the model can be further constrained by fixing $C$ to 1. This allows some comparability of experiments and expresses the rate of forgetting in terms of the
Dilution as a model of forgetting

number of null traces introduced into the trace population, $D$, per unit time as a fraction of $C$; evaluated by $k/2\Theta$. We return to this issue in the section *How do we interpret the model parameters?* at the end of this article. There we consider whether the estimates of $k$ and other free parameters of the PD model derived from the different experiments analysed here, collectively offer further insight into the processes of forgetting and the validity of the PD model.

*Applications of the PD model: long-term cued recall*

A key element in the population dilution model as articulated in Equation [1] above is that the population of traces defined as potentially accessible by a specific cue have equal status: the probability of access of any trace is strictly the reciprocal of the population size. No distinction is made between $C$, $E$, or null traces as valid recalls, and while temporal discrimination plays a part in defining membership of the population, it does not make any traces within it more accessible than others as would be the case, for example, when considering recency effects in short term memory. On the same basis, we also assume that maintenance rehearsal has not been used to keep specific items more available than others. For the purposes of examining Equation [1] as a model for the forgetting function, more useful variables are: the degree of initial learning (which will principally influence the parameter $C$); the likelihood that the experiment engenders errors of association (which influences $E$); and delay over different and widely spaced testing periods (which impacts upon $kt$). Recall experiments that usefully manipulate these variables are, in chronological order: Krueger (1929); Postman and Riley (1959); Slamecka and McElree (1983); Runquist (1983) and Rubin, Hinton and Wenzel (1999). We present below an application of the PD model to these experiments separately; each with a modest degree of adaptation to
apply the model to the specific circumstances of the experiment. Rubin et al.'s dataset, being the most substantive, is subjected to a more detailed analysis.

Krueger (1929). Three groups of participants (which we label $G_0$, $G_{50}$, $G_{100}$) learned lists of 12 words by the anticipation method at a rate of 2 seconds per word. Training to a criterion of one correct anticipation of the list generally took between 4 and 5 presentations. One third of the participants, $G_0$, ceased training upon reaching criterion, another third, $G_{50}$, were overtrained by a further 50% of presentations and the final third, $G_{100}$, by an additional 100% (i.e., a further 4 or 5 trials on average). Within these groups separate subsets were retested at 1, 2, 4, 7, 14 and 28 days both in terms of their correct anticipations on the first relearning trial (effectively a recall score) and a savings score which we consider in a later section. A complication in modelling these data is that the participants varied considerably in their speed of learning. With a mean number of trials for $G_0$ of approximately 4 and a standard deviation of about 1.3, it can readily be seen that many participants completed a correct anticipation within 3 trials whilst others took twice as long. Since overlearning was defined in terms of the number of trials in which participants met the criterion of one completely accurate anticipation, this renders the number of overlearning trials – clearly central to the PD model implementation – as variable within the same groups and the model can therefore only be approximated. Furthermore, Krueger actually only tested anticipatory responses on alternative presentations of the lists (whether starting with the first or second trial is not specified). This makes our ability to estimate the memory populations even less precise, with the impact being felt most strongly on group $G_0$ because the proportional difference between the fastest and slowest learners is greatest when there is no overlearning.
Nevertheless, within these limitations, Krueger’s experiment excludes short term rehearsal effects and allows an application of Equation [1], above, as follows. First, we assign an arbitrary weight to the constant $C$ of 1, defining the number of correct traces in group $G_0$ relative to the other groups. At time $t$ the probability of recall for each group is therefore given by three functions manipulating the constants $E, k, \Delta_{50}$ and $\Delta_{100}$, where the latter constants reflect the additional correct traces resulting from overlearning:

\[
p(R)_{G_0} = \frac{1}{1 + E + kt}
\]

\[
p(R)_{G_{50}} = \frac{1 + \Delta_{50}}{1 + \Delta_{50} + E + kt}
\]

\[
p(R)_{G_{100}} = \frac{1 + \Delta_{100}}{1 + \Delta_{100} + E + kt}
\]

Since training took the same time to criterion, the number of anticipations in error is a constant $E$ (expressed here relative to $C = 1$) and the model assumes the population dilution occurs at the same rate for each set of participants. If we assume that Krueger’s participants learned roughly a constant number of words per cycle, it happens that, on average, the number of error anticipations up to and including the first totally correct trial approximates the number of correct responses. This allows us to model $E$ in the equations above as 1 (i.e., equal to $C$), leaving all predictions for Krueger’s data mediated only by the free parameters $k, \Delta_{50}$ and $\Delta_{100}$. We illustrate the best fit of this constrained model in Figure 3 ($k = 4.695$, $\Delta_{50} = 3.705$, $\Delta_{100} = 5.574$). In this, and the experiments reported below, data are presented on natural logarithmic
scales and the model is optimised by minimising the sum of squares between model and predictions in this log-log transformation. The model accounts for 94.9% of the variance in the data. Beyond this, we do not attempt to evaluate the closeness of the fit of this model for these data or the data sets following; leaving these plots to speak for themselves.

Group G₀ show a poorer fit to the model than do the overlearning groups G₅₀ and G₁₀₀, for which some additional comment is required. In mixed population models, the performance of the smallest memory populations (i.e., those with least overlearning) will intrinsically produce greater variance in response. As we have noted, the variability between participants and the use of anticipations on alternate trials further adds to the uncertainty in the data. Also, the optimised parameter values for Δ₅₀ and Δ₁₀₀ of 3.705 and 5.574, estimating the increased number of C traces in overlearning, exceed the values of 1.5 and 2.0 that would have been expected relative to the fixed value of 1 for Group G₀. In model terms, this implies that the impact of the overlearning trials upon the value of C is greater than the original learning trials. This finding is a recurrent theme in the following analyses and an issue to which we return later. Note also, with anticipations being elicited on alternate trials only (whether odd or even trials is unknown), the criterion of one correct anticipation possibly misrepresents the true state of memory in Group G₀. Given these uncertainties in interpreting Krueger’s experiment, uncertainties that will be felt most acutely in the G₀ data, we are happy to present this fit as reasonable.

Postman and Riley (1959). Participants learned CVC lists in an A-B, A-C retroactive interference paradigm varying both the number of learning trials for list A-
Dilution as a model of forgetting

B and the number of intervening trials on list A-C before retesting list A-B. The learning conditions were varied between 5, 10, 20 and 40 trials and retention intervals were varied between 0, 5, 10, 20 and 40 trials. For the purposes of model fitting we exclude the first of these retention intervals because they are likely to be influenced by short term memory processes, although these data are included in Figure 4. In addition, the 5-learning trial condition produced detectable recall only on immediate testing and is therefore excluded.

Insert Figure 4 about here

The implementation of the PD model in this case, illustrated in Figure 4, is as follows. First, an attempt to fit the model with the assumption that each learning trial produces one correct trace (i.e., $C$ is fixed at 10, 20, or 40 for each group) provides a poor fit in which recall is overestimated. Estimating $C$ for each group as a free parameter of the data yielded values of 2.3, 12.8 and 33.1 respectively with an estimated (common) $E$ term of 9.2. This is the same pattern of disproportionality that is observed in Krueger's data, with overlearning trials apparently having greater impact upon the population of $C$ traces than earlier learning trials. To constrain the implementation of the model as far as possible, we assume here that the first $m$ trials of learning produce no $C$ traces. Whether the resulting traces are defined as null or erroneous is immaterial because they have no impact upon the probability of a correct response. This means we can effectively model $E = m$ in Equation [1] to model this data by the function: $p(R)_{CI} = (l - m) / (l + kt)$; where $l$ corresponds to 10, 20, or 40 learning trials depending upon the condition. We can then implement the PD model as in Equation [1] with just two free parameters for the entire dataset: $m$ and the
Dilution as a model of forgetting

forgetting parameter $k$. Minimising the sum of squares of the log-log transformed data (illustrated in Figure 4) gives $k = 0.92$ (delay in this case measured in terms of intervening learning trials on list 2) and $m = 7.38$; with this model accounting for 91.4% of the variance. As in Krueger’s data, the fit of these data is again related to the putative size of the memory population in each group, with the fewest overlearning trials producing the greatest variance. Note also that the data show higher than predicted recall at the shortest lag in all three groups, suggesting some short term recall, as expected.

*Slamecka and McElree (1983).* Experiment 3 in Slamecka and McElree (1983) is a commonly-cited study of the relationship between levels of learning and forgetting rates. Here, participants learned lists of 16 sentences in which the subject clauses were used as cues for recall of the predicates. In a ‘high’ learning condition the sentences were seen four times and in the ‘low’ learning condition three times. Testing occurred either immediately, after 1 day, or after 5 days where participants were supplied with sheets containing the 16 subject clauses of the stimuli with space next to them to as cues for the recall of the predicate phrase. With only two degrees of learning and three data points to fit per group, the discrimination between models is slight and we model these data for illustrative purposes only, seeking to constrain the model as far as possible. For this reason, and given participants were not given the opportunity to anticipate or articulate errors, we assume $E = 0$. In that case, the mean levels of recall of 9.3 and 11.8 at immediate recall imply that the trial sequences have not, in fact resulted in all 16 stimuli having been learned (else these values should both be 16). Following Rock (1957) and others, we assume the participants learn a fixed number of new sentences on each presentation cycle, with sentences learned on
Dilution as a model of forgetting

previous trials effectively receiving overlearning trials. This leads to some being heavily overlearned and others not learned at all (with potentially all intermediate states represented). Modelling the parameter $n$, the constant number of stimuli acquired upon each trial, and optimising that against the observed values gives $n = 3.006$. We therefore approximate this for convenience as $n = 3$. This allows us, for an average participant, to assign the number of stimuli learned on the first, second, third, and in one group, fourth learning trial. If we assume that each learning trial and each subsequent presentation of that stimulus increments $C$, the number of correct traces, we can assign a probability of recall for each stimulus moderated only by the forgetting constant $k$. For example, for the first stimulus learned on presentation 1, the probability of recall on day $t$ is given as $3/(3 + kt)$ for the low learning group and $4/(4 + kt)$ for the high learning group, bearing in mind that $E$ is fixed at zero. Determining these values and then aggregating them across all 16 stimuli therefore allows us to predict overall recall levels at each delay. These are shown in Figure 5, with $k$ optimised at 1.24. In this case, the PD model accounts for 95.3% of the variance in the data.

Insert Figure 5 about here

As noted, the near-immediate testing at the shortest delay, coupled with only two other delays means this is a weak illustration of the PD model and may explain why it is possible to model with only one free parameter. It is also impossible with this number of data points to assess whether short term effects are evident at the shortest delays. We note also that it is an arguable point whether recall in this case is actually cued recall (as we model) or a form of free recall in which participants make repeated samplings of memory to complete all 16 cues. In the latter case, it would be
surprising to see Equation [1] provide such a good account of this data, since our model assumes that the probability of recall failure is predictable from the likelihood of accessing a null trace from a single, not repeated, sampling of the population. We have assumed that given 5 minutes to complete 16 cued recall tests, that is, 20 seconds per stimulus (Slamecka and McElree do not report whether any participants requested longer), repeated sampling of memory was rare. We include this reanalysis of Slamecka and McElree’s Experiment 3 despite these concerns because it is a much-discussed study in respect to modelling the forgetting function and it is useful to show that the PD model can fit these data with ease.

Runquist (1983; Experiment 2). In this study, participants saw 24 word pairs as stimuli and were tested by cued recall of the second word by the first at delays of 20 minutes, 1 and 6 hours, and 2, 7 or 21 days. Half of the participants were shown the stimuli just once, and the others saw them repeated on three occasions. Additionally, after a 2 minute distraction task, all the participants were tested for recall for half of the stimuli. This allows us to consider recall across the retention intervals as a function of six contingencies: for untested stimuli whether the stimulus was seen repeatedly or just once (referred to as U3 and U1 below); whether it was tested and recalled correctly for repeated and non-repeated stimuli (TC1 and TC3); and whether it was tested and not recalled for repeated and non-repeated stimuli (TE1 and TE3).

We model these six contingencies with five free parameters as follows. First, as before, we fix the number of correct traces, $C$, generated from a single stimulus presentation at 1. The parameter $k$ models the influx of null traces into the trace
Dilution as a model of forgetting

population, $C_3$ models the number of traces supporting correct recall derived from three repetitions of the stimulus, and $E$ models a fixed number of error traces; as before. To account for the influence of the immediate testing procedure on half the stimuli, $C_{\text{test}}$ models the number of correct traces added to the trace population by dint of a correct recall, and $E_{\text{test}}$ increments the number of error or null traces that are generated following incorrect recall at this stage. These could occur, for example, if an error was uttered that is subsequently remembered, or if the participant encodes a failure to recall as a null trace.

Implementing these parameters into Equation [1] gives the following retention functions for each of the experimental contingencies:

\[
p(U1|t) = \frac{1}{1 + E + kt}
\]

\[
p(U3|t) = \frac{C_3}{C_3 + E + kt}
\]

\[
p(TC1|t) = \frac{1 + C_{\text{test}}}{1 + C_{\text{test}} + E + kt}
\]

\[
p(TC3|t) = \frac{C_3 + C_{\text{test}}}{C_3 + C_{\text{test}} + E + kt}
\]

\[
p(TE1|t) = \frac{1}{1 + E + E_{\text{test}} + kt}
\]

\[
p(TE3|t) = \frac{C_3}{C_3 + E + E_{\text{test}} + kt}
\]

These functions optimise the collective fit of the data with values: $C_3 = 2.115$, $E = 0.422$, $E_{\text{test}} = 9.478$, $C_{\text{test}} = 9.646$ and $k = 0.0429$; with 96.7% of the variance accounted for thereby. As a matter of note, the additional constraint that $E_{\text{test}}$ equals $C_{\text{test}}$ (clearly possible from the estimated values), leaves that variance and the other optimised parameter values effectively unchanged, and it is this constrained version of the model that is illustrated in Figure 6 alongside Runquist’s data.
We discuss the interpretation of the optimised parameter values in the General Discussion (below). For the present, it is sufficient to note that the PD model gives a good account not only of the overall data, but it also (applying the same forgetting mechanism in each case) predicts retention curves for those stimuli recalled correctly on the initial test (where subsequent retention is significantly increased) and for those not recalled correctly on the initial test (where recall is very much lower as a result). Note also that the presentation of the data in a log-log transformation in Figure 6 exaggerates the absolute discrepancy between model and data for low frequencies.

Rubin, Hinton and Wenzel (1999). Rubin et al. reported both a recall and recognition experiment. This section presents an analysis of the recall data only, with the recognition experiment considered in a later section. Their study was conceived with the specific purpose of generating a dataset sufficiently precise to enable discriminating comparisons of different retention functions; previous research (e.g., Rubin & Wenzel, 1996) having established that most datasets cannot. In their recall experiment, 300 participants saw word pairs in a continuous presentation per trial procedure. Recall for the second word of the pair was cued with the first after lags of 0, 1, 2, 4, 7, 12, 21, 35, 59, and 99 trials. The experiment therefore spans both short term and longer term recall. In the analysis below, we present the full range of data but fit our model only to data of lag 4 or longer; consistent with our strategy of excluding short term processes. Following Rubin et al. (1999, Figure 4), we classify the 300 participants into quintiles to model the retention function for five groups segregated according to their degree of initial learning.
The division boundaries for these quintiles were .265, .315, .380, and .465, corresponding to mean overall performance of .230, .268, .343, .414, and .577 respectively in the separated groups. Equation [1] can then be applied to the data from each of these groups with \( E \) and \( k \) as common parameters. We assume an error term is required because real word pairs were used. This is equivalent to saying that \( E \) incorrect associations were available to a given cue at lag 0 because of extra-experimental proactive interference. In these model fits \( E \) is estimated (relative to \( C = 1 \)) as 4.145 and the forgetting parameter \( k \) at .293. Values of \( C_n \) were optimised as free parameters for the four highest quintiles (the lowest being arbitrarily fixed at 1). The resulting estimates of 1.768, 2.622, 4.260, and 9.968 respectively correspond to the patterns seen in previous model fits in which the higher performing groups estimate values of \( C_n \) are disproportionate to the relative value of 1 fixed for the lowest performing group\(^7\). This PD model accounts for 99.1% of the variance. For the purposes of fitting these data, however, the estimated values of \( C_n \) admit of further constraint because they very closely approximate to a quadratic relationship between mean levels of performance and estimated values of \( C_n \) according to the equation \( C_n = \alpha R_n^2 - \beta \); where \( R_n \) is the mean overall performance for a given quintile, and \( \alpha \) and \( \beta \) are constants optimised at 2907.1 and -52.4 respectively. For parsimony only, we use this relationship to predict \( C_n \) values for all five groups but make no theoretical claims for the quadratic form of this function.

Insert Figure 7 about here
Dilution as a model of forgetting

The optimisation of the PD model for the Rubin et al. data therefore employs 4 free parameters and is optimised against the 35 data points in which the lag between presentation and test is equal or greater than 4. This model accounts for 99.0% of the variance. Predictions for this version of the PD model applied to Rubin et al.’s data are shown in Figure 7 (reproducing Rubin et al.’s Figure 4 with the PD model predictions superimposed as continuous lines). For note, values of $C$ are estimated as 1.017, 1.886, 2.903, 4.457, and 9.145 for the lowest to highest performing quintiles respectively. $E$ is estimated at 4.351 (implying a high level of extra-experimental interference) and the forgetting constant $k$ at .303. The key point here to which we draw attention is the ability of the PD model to predict the shape of the retention function, as illustrated in log-log transformed plots. The model also does an excellent job of predicting these functions (both in shape and absolute levels of performance) for each of the five quintiles of participants representing different levels of initial learning. Disregarding the shortest lags that clearly show the effects of short term recall upon memory (note also that the shortest lag of 0 is excluded from Figure 7), the PD model is particularly successful in modelling the changing gradient of the retention function with increased delay; a characteristic noted by Rubin et al., and which Appendix A demonstrates is a logical consequence of the PD model as expressed in Equation [1].

Applying the PD model to recognition and savings
Since Luh (1922), there has been a general understanding that other measures of retention – specifically recognition and savings – produce different forgetting functions to those seen in cued recall. Recognition generally produces much higher levels of performance and forgetting curves derived from savings measures generally show very rapid forgetting followed by apparently much slower decline; allowing an
Dilution as a model of forgetting

approximately linear function of savings modelled in log time (Ebbinghaus, 1885; Luh, 1922; Krueger, 1929; see also Woodworth & Schlosberg, 1954, Figure 23.15). Does the PD model apply to recall only or can it deliver a generic approach to modelling these other forgetting functions?

In this section we consider four recognition memory studies, covering a range of methodologies, that have a different impact upon how the PD model is implemented. These culminate in the presentation of a more general version of the PD model – referred to as the PDE (Population Dilution Extended) model – to which the earlier version embodied in Equation [1] is, in fact, an approximation. First, we reanalyse Shepard’s (1967) experiment on the recognition of pictures. This is a much-quoted study that allows us to introduce some principles of Signal Detection Theory. It is therefore a useful starting point in applying the PD model to recognition data. Second, we present an analysis of Strong’s (1913) experiment. This experiment is interesting here because the forgetting function reported has been convincingly plotted as a linear relationship between percent recognised and log time elapsed (see Woodworth & Schlosberg, 1954, Figure 23.14); a property inconsistent with the PD model. Reconciling this apparent inconsistency allows us to make a number of other observations about modelling the forgetting function. Third, we re-analyse Fajnsztejn-Pollack’s (1973) study. This is particularly appropriate for our purposes for three reasons. First, it investigates memory over very widely spaced intervals (2 to 49 weeks). Second, it manipulates the number of presentations of the stimuli to allow a more robust test of the PD model. Third, the old-new paradigm used, and specifically an observed change of false alarm rates with increasing delay, requires a full specification of the PDE model to account for the data over the full time range.
and variations of learning conditions. The fourth experiment presented in this section is Rubin et al.'s (1999) recognition experiment. Following the presentation of recognition experiments, we end by considering the applicability of our approach to savings measures of forgetting.

Applications of the PD model: recognition.

The PD model for recognition proposes the same processes as in the model for cued-recall; illustrated in Figure 2. Central to these is the idea that stimuli have multiple representations in memory and retrieval is a process of selecting and using just one. First, a segment of memories within the mnemonic record is identified according to the principles of temporal organisation of memory. Second, as before, we assume a filtering of that segment that identifies a subset of traces whose strength of association to the present cues for recognition exceed a threshold value. These traces comprise the population of accessible traces which allow us to predict the outcome of the recognition process. As before, because the PD model relates forgetting in this population to the influx of additional traces relative to the size of that population, this threshold is not specified. Third, a sampling of a single trace from that population is used as a basis of response, depending upon the recognition methodology used. Here, however, an important difference arises between models of recognition and cued recall. In cued recall, random sampling from the population allows the participant to use the outcome as a response (correct or incorrect) or accept that no response is possible when the recalled material is null or readily censored. Recognition, on the other hand, is more constrained because the participant must make judgements on the recognition cues rather than making a response. The simplest approach, following from the PD model used in cued-recall experiments, would have been to assume that a
Dilution as a model of forgetting

retrieved trace is matched against the recognition cues in an all-or-none way. Thus, in an old-new paradigm, a match of recognition cue and a retrieved trace allows an ‘old’ response and a non-match a ‘new’ response. In forced-choice paradigms, a match would allow selection of the matching alternative response and a failure to match would be followed by a simple guess between the two alternatives. However, this high threshold theory (e.g., see Macmillan, 2002) is now seen at best as an approximation of memory performance because recognition is rarely all-or-none in practice. For example, working in a framework of Signal Detection Theory (SDT), Wixted (2007) claims that every ROC curve derived from recognition memory experiments analysed between 1958 and 1997 was curvilinear; a property incompatible with high threshold theory. It has therefore become common practice to apply the assumptions of SDT to the modelling of recognition.

The detailed terms under which this is done are still the subject of considerable debate (e.g., see Glanzer, Kim, Hilford & Adams, 1999; Yonelinas, 1999; and, Wixted 2007b; Parks & Yonelinas 2007). Here, consistent with our aim of evaluating the generality of population dilution as a model of forgetting, we develop the PD model with elements of SDT to accommodate the challenges recognition data present. However, the primary aim is to remain as consistent as possible with the basic assumptions and characteristics of the PD model as laid out in the introduction to this paper. This results in an application of SDT that differs from those commonly used to model recognition performance. The key element in this is that, whereas in standard approaches decisions are commonly modelled by the subtraction of one strength value from another (as in between the strength of a target and foil), we are interested in larger populations of traces of varying strength, and use decisions based upon
Dilution as a model of forgetting

associative strength to determine the proportion of those populations actually eligible for retrieval. In other words, criterion strength values in the PD model are used to manipulate the composition and numerosity of trace populations from which random selection determines the probability of correct performance. The simplest paradigm within which to demonstrate this is Shepard’s (1967) two-alternative forced-choice recognition experiment (2AFC); with which we begin.

Shepard (1967). Shepard showed participants 612 complex pictures chosen to be individually of high distinctiveness and collectively of low confusability. Recognition was tested immediately and for different subsets of participants after 2 hours or 3, 7, or 120 days with a 2AFC procedure. We implement a version of the PD model to predict the likelihood of identifying the correct alternative as a function of time on the following principles. First, temporal discrimination identifies a segment of the memory record of length $2\Theta T$ that is centred upon the stimulus presentation period. Within this segment, traces accessed by both the target and foil cues exist with differing strengths of association with the recognition cues and with different densities per unit time in the memory record. As discussed in the modelling of cued recall, we are interested in this stage in only those memory traces that meet a minimal threshold of association with the available prompts for retrieval; the level of which need not be specified. Thus, we assume for the $p$ seconds when the stimuli were actually presented, the density of accessible traces encoding the stimuli is relatively high at $s$ traces per second; accessing $sp$ traces in all. This is constant and not a function of time because the segment of width $2\Theta T$ is assumed to include the entire duration of the stimulus. Over the $2\Theta T$ seconds of the memory record, a density of $f$ traces per unit time accessible to the foil gives rise to $2f\Theta T$ traces. Furthermore, the
Dilution as a model of forgetting

target cue acts in the same way as the foil in overlapping with traces in the accessed memory record other than those originating from the stimulus presentation. Assuming the stimuli and foils are matched in saliency and frequency of use, it follows that the number of these traces equals the $2f \Theta T$ traces activated by the foil. Thus the trace population in the PD model comprises $sp + 2f \Theta T$ traces available to the stimulus item and $2f \Theta T$ traces available to the foil item.

Adopting a signal detection perspective, we now assume that this population varies in the degree of overlap with the recognition cues that can be treated as synonymous with strength. For those $2f \Theta T$ traces accessed by the foil, the probability distribution of strength within the population is described, by convention, by a Normal distribution of mean strength 0 and standard distribution 1; denoted $N(0,1)$. Traces activated by the stimulus item are described by two distributions. The strength of those $sp$ traces derived from the stimulus presentation itself are described by $N(m,\sigma)$, a distribution of mean $m$ and standard distribution $\sigma$, representing a greater mean strength and a variable distribution. Additionally, by the above logic, a second distribution of $2f \Theta T$ traces strengths exactly matches that of the foil because the stimulus item also activates traces not originating from the stimulus; with a strength distribution of mean 0 and standard distribution 1. The distribution of trace strengths activated by the target recognition cue is therefore modelled as the sum of two Gaussian distributions, as illustrated in Figure 8.

Insert Figure 8 about here
Dilution as a model of forgetting

We now assume that the trace population is further reduced by a decision process which excludes from the population traces that fall beneath a criterion strength $\lambda$ (also shown in Figure 8). This decision process is additional to, and separate from, the filtering process in the PD model in which traces are activated by the recognition or recall cues available. The proportion, $P_n$, of non-stimulus traces activated by the foil is therefore given by $\int_0^\infty N(0,1) \, ds$; as is the proportion of non-stimulus traces activated by the target. The proportion, $P_s$, of target traces activated by the target cue is $\int_0^\infty N(m,\sigma) \, ds$. Provided $\lambda$, $m$ and $\sigma$ are constant, these proportions are constant in a given experiment. In the absence of any evidence to the contrary, we assume this to be so for Shepard’s experiment – although, as will be seen, this is not universally the case. Thus the traces activated by the recognition cues are $P_{sp} + 2P_{sf/\Theta T}$ from the stimulus item (selection of any of which will elicit the stimulus as the response choice) and $2P_{sf/\Theta T}$ for the foil, whose selection would result in an error. For this 2AFC paradigm, the PD model therefore gives a probability of a correct selection as a function of time $t$ as:

$$P_{correct} = \frac{(1 + kt)}{(1 + 2kt)} \quad \text{where the constant } k = \frac{2P_{sf/\Theta T}}{P_{sp}} \quad [3];$$

that shares the same logic as the PD model for cued recall expressed in Equation [1]; differing only to take account of the different testing circumstances. In Shepard’s experiment, correct recognition varied from 99.7% accuracy after two hours to 57.7% after 120 days, and Figure 9 illustrates the fit of this model to the data (optimised by minimising the sum of squares, represented, as above, in log-log form), with the single free parameter $k$ optimised at 0.079. The model provides a very close fit to the data, accounting for 99.8% of the variance.
Dilution as a model of forgetting

Strong (1913). Strong presented 20 words to participants and subsequently tested recognition of these words when mixed in a list of 20 lures over 13 different delays ranging from immediately to 7 days. Participants were required to indicate which 20 of the 40 given words were stimuli and to assign a confidence to each selected item from 1 (sure) to 4 (guess). We are interested here in the proportion of targets correctly identified regardless of confidence level. This varied from .92 to .57 at the shortest and longest delay respectively. In fact, interpreting the reported data is problematic for a number of reasons, including the fact that the reported forgetting function is predicated upon an arbitrary weighting of responses as a function of the confidence levels given. Also, the participants were required to offer recognition judgements on exactly 20 of the 40 words offered only, essentially constraining the experiment to a forced-choice recognition task; albeit a complex one. Nevertheless, the detail with which Strong reports his experiment makes it possible to re-analyse his data from the raw scores presented in his paper to estimate a proportion $P$ of correctly identified targets, irrespective of the confidence level assigned to them. We implement the PD model in the same way as for Shepard’s experiment by assuming, for the sake of analytical convenience, that the experiment can be treated as comparable to 20 2AFC tasks.$^9$

The probability of any word being correctly recognised as a function of time is given, as before, by Equation [3] above with the additional constraint (following a remark by Strong in the original article) that some words were noticeably harder to learn than others, leading to a probability $L (<1)$, that the item will survive in memory beyond the very short term. In the event that an item is not retained, both target and
foil cues are of equivalent strength and the probability of a correct selection is 0.5, leading to the modified form of Equation [3]: $P_{correct} = L(1 + kt)/(1 + 2kt) + (1 - L)/2$. The briefest delays in this experiment admit of the possibility of short term effects in recognition, and to enable us to model across the entire range of delays in this case, we incorporate an approximation to model the availability of stimulus material in short term memory\(^1\) of the form $P_{stm} = e^{-\alpha t}$, where $\alpha$ is a free parameter of the model. If available, this short term representation is assumed to take precedence in retrieval, and the predicted proportion of stimuli recognised as a function of time is therefore given by $P_{correct} = e^{-\alpha t} + (1 - e^{-\alpha t}) (L(1 + kt)/(1 + 2kt) + (1 - L)/2)$. This theoretical function, optimised for delays of 30 minutes and longer (i.e., the short term data was not used to optimise the model) with parameter values $L = .541$, $\alpha = .155$, and $k = .544$. This function is shown in Figure 10 as the continuous line. The dotted line represents the best fit of a log-linear function of the kind $P_{correct}|t = m\log(t) + c$; where $m = -0.031$, $c = .896$. There is little to choose between the two models; the log-linear model accounting for 92.1% of the variance, compared to the PD model accounting for 91.3%. There is certainly no basis to exclude the PD model as a reasonable account of this data. Moreover, our secondary purpose was to demonstrate that the perceived linearity in Figure 10, ostensibly contradictory to the PD model, is readily approximated by it across the entire range of delays, provided a rudimentary account is made of the added effect of short term memory for short delays.

_Fajnsztejn-Pollack (1973)_ Fajnsztejn-Pollack presented a group of children and adolescents a set of 280 pictures once, twice, or four times in a learning session and
Dilution as a model of forgetting

subsequently tested recognition at delays of 2, 5, 10, 20, and 49 weeks. The data is useful for our purposes both in evaluating the impact of repeated stimulus presentations upon recognition memory and because it covers a wide range of delays. In developing the PD model with Signal Detection Theory to include recognition performance, this experiment is also important because the old-new paradigm constrains the application of the PDE model in interesting ways; including an observed shift in the rates of false alarms with delay that is of considerable theoretical significance to the modelling of the forgetting function.

Recognition was tested by presenting a sequence of 84 pictures of which 28 were new, 28 had been seen once, and 28 had been seen repeatedly (two or four times, depending upon the group of participants). The participants were required to indicate whether each picture was new or old. For the present purposes we analyse the data from adolescents (15-16 years old) only. Participants were in two groups, A and B (our labelling). Group A saw pictures either once or twice and Group B saw pictures once or four times. The four rows of data therefore represent data collected from two sets of participants. Fajnsztejn-Pollack’s data for this set are reproduced here in Table 1 in the form of correct hit and false alarm proportions for each delay and training contingency.

We assume, as in the analysis of the Shepard and Strong data above, that all traces accessed by the recognition cues vary in strength. As before, irrelevant traces

37
Dilution as a model of forgetting

are assumed to be of mean strength 0 with a standard deviation of 1, and the traces
derived from stimulus presentations to be of mean strength \( m \) and standard deviation
\( \sigma \), both free parameters of the model. In this experiment, participants saw either a
new picture or a stimulus picture, and were asked to make an old-new decision. It is
useful to consider first the processes assumed to apply when the recognition cue is
new, where we need the model to predict the likelihood of an ‘old’ response (i.e., a
false alarm). This requirement is modelled by the insertion of the additional decision
process described first in the analysis of Shepard’s data and illustrated in Figure 8.
Thus the population of traces activated by the new picture are subject to a decision
process in which the proportion of traces exceeding a criterion strength \( \lambda \) determines
the likelihood of making an ‘old’ response. The likelihood of a false alarm is given
by the proportion \( \int_{-\infty}^{\infty} N(0,1) \, ds \) of null traces exceeding the criterion value
\(^{11}\).

In the case that an old item is shown, the situation is more complex because
calculating the proportion of traces of strength greater than \( \lambda \) applies to the
combination of two differently weighted distributions – one arising from irrelevant
traces that happen to be activated by the target, and the other from the original
stimulus presentations with which it is more strongly associated. Furthermore, we are
considering not only the distribution of strength of these traces, but also their
numerosity. We begin by assuming as before that the number of correct traces arising
from a single presentation of a stimulus is 1. We can then model the number of traces
arising from 2 or 4 presentations with two free parameters, \( R_2 \) and \( R_4 \). Relative to this,
the number of irrelevant or null traces activated by an old item as a recognition cue as
a function of delay \( t \) is \( kt \), where \( k = \frac{2f \Theta}{sp} \). The overall population size of traces
Dilution as a model of forgetting

relevant to the judgement of old stimuli is therefore \((R_n + kt)\), where \(n = 1, 2 \text{ or } 4\) and \(R_1 = 1\). \(R_2\) and \(R_4\) are free parameters modelling the effect of repeated presentations of the stimulus, and \(k\) models the relative balance of null and stimulus traces per unit time in the memory record.

To remain consistent with the mechanism for predicting the level of false alarms, we again assume that the population of eligible traces is subject to a decision process in which only traces exceeding a criterion value \(\lambda\) are designated as ‘old’\(^{12}\). For stimulus traces, this proportion \(P_s = \int_0^\infty N(m,\sigma) \, ds\), and for irrelevant traces, the proportion \(P_n = \int_0^\infty N(0,1) \, ds\). This gives the probability of correctly recognising a stimulus item in an old-new paradigm as:

\[
P(\text{’old’}|\text{old}) = \frac{k t P_n / (R_n + kt) + R_n P_s / (R_n + kt)}{R_n + (R_n + kt)}
\]

[4]. Finally, we note that the false alarm rate in this experiment increases with delay, which we interpret as a decline in the decision criterion \(\lambda\) as a function of time (see also Gardiner & Java, 1991; Singer & Wixted, 2006). Ignoring this trend by assuming a constant \(\lambda\), or treating \(\lambda\) as a linear function of delay produces poor fits of the overall model to the data. As a matter of practice, the function \(\lambda|t = a(1+bt)^{-1} + c\), where \(a\), \(b\), and \(c\) are constants, is used here; chosen because of its close relation to the PD model as expressed in Equation [1] and the presumption that criterion shifts will be related to aspects of memory performance. This function therefore relates \(\lambda\) to overall memory performance with a maximum value at \(t = 0\) declining to an asymptotic value \(c\) at long delays. We return to consider the form of this function, and the relationship between memory performance and \(\lambda\) in the final sections of this paper. Overall, therefore, this model employs eight free parameters to model
Dilution as a model of forgetting

Fajnsztejn-Pollack’s data. The optimised fit of this model is illustrated in Figure 11. $k$ is estimated at 0.466; which translates, with appropriate adjustments, to an influx of 0.676 null traces per day relative to $C_1$ fixed at 1. $C_2$ and $C_4$ are estimated at 5.877 and 42.24 respectively. The strength of stimulus traces relative to the noise distribution $N(0,1)$ is estimated by a normal distribution of mean 4.987 and standard distribution 1.435, and the parameters $a$, $b$, and $c$, modelling the changing criterion level $\lambda$, are optimised at 5055.40, 1465.89, and 0.78. These latter estimates generate values of $\lambda$ of 4.23 after one week; declining to 0.86 after 49 weeks. Overall, the fit to this model (accounting for 97.2% of the variance) is good for the 30 data points modelled, and we note for future reference that the relative weights for repeated presentations of the stimulus, $R_2$ and $R_4$ are once again proportionally large in relation to the assumption of $R_1$ being equal to 1.

As a final point, we comment on the common observation that 2AFC performance is generally higher than that in old-new paradigms. In terms of standard signal detection models, $d'$ for 2AFC experiments is predicted to be $\sqrt{2}$ of that observed in old/new experiments; although that is predicated upon assumptions about the distribution of trace strengths that do not apply in the PDE model. Nevertheless, it is possible to show that the general advantage of 2AFC experiments applies. Equation 3, giving the levels of correct performance in the 2AFC experiment can be re-written as:

$$p(\text{correct}|\text{2AFC}) = \frac{(S + Nxt)}{(S + 2Nxt)};$$

where $N$ and $S$ are the proportions of stimulus and null traces reaching a criterion value of strength $\lambda$, $t$ is time elapsed, and $x$ is the positive ratio $2f\Theta/sp$ reflecting the relative ratio of interfering traces to stimulus
Dilution as a model of forgetting

encodings for unit time in the memory record. Equation 4 models the likelihood of correctly identifying a stimulus in an old-new paradigm. Using the same terminology, and assuming: i) that there are no stimulus repetitions (for comparability with Equation 3); and, ii) accounting for the fact that the overall probability of correctness includes the correct identification of new stimuli; gives $p(\text{correct}|\text{old-new}) = \frac{(S + N_{xt})}{2(1 + xt)} + \frac{(N - 1)}{2}$; assuming old and new targets are equally likely in the experiment. The difference, $\Delta$ between the two, giving the advantage of recognition in the 2AFC paradigm compared to the old-new paradigm, is therefore given by:

$$\Delta = (S + N_{xt}) \{ \frac{1}{S + 2N_{xt}} - \frac{1}{2 + 2xt} \} + \frac{(1 - N)}{2}$$

Given that $S$ and $N$ are proportions between 0 and 1 and therefore $(S + 2N_{xt}) < (2 + 2xt)$ it can be seen that $\Delta$ is necessarily positive, although the degree of difference depends upon the values of $S$, $N$ and $x$. We argue in the General Discussion that these are probably determined by specific experimental circumstances and therefore, within the framework of the PDE model, there is no strong reason to suppose the exact advantage of the 2AFC paradigm over the old/new paradigm can be quantified on theoretical grounds alone.

Rubin, Hinton and Wenzel (1999). Rubin, Hinton and Wenzel’s (1999) recognition experiment employed a continuous recognition paradigm in which stimuli (digit-letter-digit trigrams in which the digits 1-9 and the consonants K, V, W, Y and Z were used) appear twice within a long sequence with differing lags (0, 1, 2, 4, 7, 12, 21, 35, 59, and 99) between the first and second presentation. On the second occasion, participants were either required to make an old-new judgement or a ‘recognise-know’ judgement. We concentrate here upon old-new judgements only; the PDE model in its present form having nothing to say about recognise/know
Dilution as a model of forgetting

judgements. For the purposes of applying the PDE model, the data are noteworthy for a high false alarm rate (estimated overall at .643) that increases systematically through the experiment, as occurred in Fajnsztejn-Pollack’s experiment.

To apply the PDE model therefore requires some interpretive analysis of the data reported in Rubin et al. First, because they apply a high threshold recognition model to convert observed recognition rates into estimations of recognition probabilities (e.g., see Table A2 in Rubin et al.), the reported recognition rates must be converted back to estimated hit rates to allow the PDE model to be applied. This requires us to estimate false alarm rates for each of the test delays separately whereas the available data (Rubin et al., Figure 10) only provide this as a function of absolute trials in the experiment; aggregating false alarm rates across different test delays. These rates must differ for different specific delays given the very sharp rise in false alarm rates throughout the experiment evident in Rubin et al.'s Figure 10 and the distribution of testing schedule. For example, by the time the first test at delays of 99 trials is possible at trial 100 of the experiment, overall false alarm rates have risen from approximately 0.35 to 0.5. To quantify this shift in false alarm rate we modelled the false alarm probabilities provided by Rubin et al. with the function \( p(FA|\text{trial}) = 1 - x(1-y\cdot\text{trial})^{-1} \), with \( x \) and \( y \) estimated at 0.6435 and 0.002 respectively. The repeated structure of the experiment (in which a given sequence of presentation and testing over different delays within 200 trials was repeated twice) then makes it possible to use this function to estimate the average false alarm rate for the longer lags 55 and 99; because tests at these lags can only occur at identifiable trials within the experiment. Similarly, the average false alarm probability for delays of zero can be assumed to equal the overall average; those trials being evenly distributed across the entire testing sequence. As it happens, these three estimates for re-test delays of 0, 55, and 99 trials
Dilution as a model of forgetting

approximate well to a linear function from which we extrapolate for all other delays. By this means, we were able to estimate false alarm rates for each delay in the experiment and convert Rubin et al.’s reported data into a form appropriate to applying the PDE model.

These interpretive analyses accepted, Equation [4] above can be applied with $R_n = 1$; there being no repetition of stimuli in this experiment. The optimised values of this function give $k = .0696$ (which, assuming crudely that the participants carry out one trial every five seconds, translates to a null-to-stimulus influx ratio of 11668.20 per day); and the mean and standard deviation of the target strengths to be 0.000 and 0.221 respectively. The fit of this model is illustrated in Figure 12. It accounts for 99.3% of the variance, and is certainly sufficient to establish the PDE model as a plausible account of these data. Additionally, the model predicts well the deviations from the power function observed at medium and long lags illustrated in Figure 8 of Rubin et al. (1999).

Insert Figure 12 about here

* Savings – Krueger (1929). In this section we aim to show how the PD model can account for the different form of the retention function typically seen with savings measures. Ebbinghaus’ (1913, Chapter 7) would seem an obvious study to model, and has the advantage of not aggregating data across participants; Ebbinghaus having used only himself. However, Ebbinghaus’ data is recorded in times to relearning, with the number of recitations unspecified. This, with a number of further adjustments to the data due to variables Ebbinghaus could not control, means that the additional assumptions required to implement the PD model would obscure, and
Dilution as a model of forgetting
detract from, the purpose of our exercise; which is to demonstrate how the PD model can apply to cued recall, recognition and savings measures (see also Waters 1941, Haupt, 2001, for comments on Ebbinghaus’ methodology). Here, we use Krueger’s (1929) study, whose cued recall data are analysed above. These data are aggregated across participants, but there are sufficient details to allow a reasonable specification of the PD model to compare to the savings data. This is critical in the application of the PD model to savings data because there are several complex processes to consider in savings methods. First, there is the process of learning itself. How, for example, does the anticipation method of learning impact upon a multiple trace model in which all anticipations (some of which will be wrong), and the feedback, are in themselves memorable? We need then to consider access of the consequent memory population as a function of time, which is the business of the PD model itself. Finally, after the first anticipation trial following a delay, which can reasonably be treated as a standard recall trial, we need a model of how further anticipations (as the participants re-learn the list to criterion) interact with the retrieval process. Given that many critical operational details of Ebbinghaus’ and Krueger’s studies are no longer available (or were never recorded), this makes the modelling of the data and the mathematical specification of a forgetting function for savings problematic.

Insert Figure 13 about here

Our approach to this here is to provide a simulation\textsuperscript{13} of Krueger’s (1929) experiment to demonstrate that how the key comparisons of savings and recall methods follow from the PD model. Figure 13 illustrates the principal elements of this comparison with the observed data in the $G_{100}$ group in Krueger (1929).
Compared to recall data, (defined as the probability of being correct on the first anticipation trial after delay), the savings measure is slightly lower at the shortest delays, but declines at a slower rate to a higher asymptotic level. The lower initial level is trivially explained by the definition of savings, which is as follows. If $I_L$ represents the number of trials initially required to achieve criterion learning (for example, one fully-correct anticipation) and $R_L$ represents the number of re-learning trials required to re-establish criterion performance after delay, proportional savings is defined by $(I_L - R_L)/I_L$. However, since the minimum $R_L$ is 1, this means the maximum savings is limited to $(I_L - 1)/I_L$, which in experiments such as Krueger’s will be about 75% (because participants typically take about 4 anticipation trials to reach criterion). The more interesting question is why the savings function remains higher at longer delays than does the recall function, and as it happens, this appears to be only indirectly related to the PD model.

The simulation is designed as follows, and based upon Krueger’s experiment in which 12 items are learned, but with anticipations tested on each learning cycle. Following Robinson and Brown (1926) and others, we assume that participants learn an approximately constant number of stimuli within the set on each anticipation trial, the proportion of the overall set being modelled as $P_L$. Thus, after the first training cycle, they will be able to correctly anticipate approximately $12P_L$ of the stimuli on the second cycle. To maintain the constant number of anticipations learned on each learning cycle, $P_L$ is incremented on each trial according to the number of stimuli $N$ learned on the previous trial such that $P_{L_{new}} = P_L/(12-N)$ while $N < 12$; at which point $P_{L_{new}} = 1$. Learning is assumed to be an absorbing state during the learning stage, so
Dilution as a model of forgetting

that once a stimulus item is learned, anticipations will always be correct until the
continuous training cycles stop.

In the PD model, every time a participant correctly anticipates a stimulus, the
population of correct traces is incremented by one. This also occurs when feedback is
given after anticipation, since this is necessarily correct. Similarly, if the anticipation
is in error, the population of error traces increases by one. Overall, therefore, this
simulation generates for each item within the list of 12 items a mnemonic record of
how many correct and incorrect representations occur over the entire training cycle
and also the number of training cycles required to reach criterion; largely determined
by the value of the learning parameter $P_L$. Note that correct anticipation in this
simulation follows from one of two routes. If the item is deemed to be in a learned
state, this occurs with probability 1. Otherwise, consistent with the PD model, correct
anticipation can still occur with a probability $C/(C+E+k_t)$; where $k_t$ is a constant. $t$
represents the number of anticipation trials already conducted in the training stage,
and $C$ and $E$ represent the number of cumulated correct and error traces in the
mnemonic record for that particular anticipation. When learning is complete to
criterion ($N = 12$) after $I_L$ trials, the simulation can be continued for a further $I_L/2$ or $I_L$
trials to simulate the 50% and 100% overlearning conditions in Krueger (1929); the
effect of which will be to increase the $C$ population for each anticipation by 2 (one for
the anticipation and one for the feedback) whilst $E$ remains constant because no
further errors are committed. After a delay $t$, recall on the first relearning trial (the
data presented and modelled in Figure 3) can therefore be predicted by Equation [1].
The relearning cycle then proceeds as before to criterion and the number of trials
Dilution as a model of forgetting

taken to achieve this, $R_L$, is used with the number of initial learning trials taken to reach criterion, $I_L$, to provide a proportional savings score.

One might expect that $R_L$ will be less than $I_L$ (producing a positive savings score) because the first relearning trial has the advantage of recovering prior memories according to the PD model. However, in earlier versions of the simulations, initially high recall at relearning gave too little advantage to be consistent with observed patterns of savings. The only device that readily matched the relatively high levels of savings compared to recall at long delays was to implement the assumption that in the relearning stage, the proportion of anticipations learned on each cycle, (initially defined by $P_L$) is greater in relearning than in the original learning cycles. This is, in fact, consistent with our analysis of Runquist’s (1983) experiment that retest trials have an inordinate impact upon memory. In our simulation, an additional parameter $P_{Lr}$ is used to model this higher probability of learning. Therefore, the simulation we present of Krueger’s data in Figure 13, compared to the observed data, is based on parameter values of $P_L = 0.3$, $P_{Lr} = 0.5$ and $k = 6$. We see no particular difficulty in the added assumption of quicker learning at retesting – this is a reasonable assumption to make for several reasons (e.g., see Nelson, 1978; and also Wickelgren, 1972) – but it is important to be clear that it is this assumption, and not fundamental elements of the PD model, that give rise to the higher savings values. Overall, the conclusion is that through simulation we can predict the relationship between recall and saving scores over time, with the PD model providing a strong prediction of the former and a contribution to the latter. However, because of the centrality of relearning in savings measures, they are not incisive measures of the retention function when studying the processes of forgetting.
General Discussion

The application of the PD model to these nine studies illustrates its explanatory power in a range of experimental circumstances and methodologies of testing. It delivers a principled explanation for regularities such as Jost’s law and has no difficulty modelling the commonly observed near-parallel retention functions with different degrees of initial learning as in, for example, Slamecka and McElree’s data; an issue that exercised earlier researchers of the retention function. The PD model can reconcile (with an appropriate account taken of different methodologies) the differences in retention functions observed in cued recall, different recognition paradigms, and savings measures. The model also predicts a negatively accelerating function when expressed in log-log transformations; a systematic deviation from the commonly-favoured power function evident in all the datasets analysed here and noted by Wixted (2004, p. 871) and Rubin et al. (1999, p. 1165). This characteristic is also perceptible in other published log-log transformations of data (e.g., see Anderson 2000, Chapter 7). The model even offers a plausible account of how Strong’s (1913) representation of the retention function as a linear relationship between recognition probability and log time is misleading. Finally, the PD model is relatively parsimonious. For example, in fitting the Rubin et al. (1999) recall data, the experiment most capable of discriminating between different models, only 4 free parameters are estimated. Overall, we believe this establishes a coherent case for population dilution as a model of forgetting.
Dilution as a model of forgetting

The remainder of the General Discussion focuses upon a review of the assumptions of the PD model and interpreting the implementation of the model to known datasets. We also relate our findings to other approaches to forgetting, since, whilst the PD model provides a good account of the retention function, it does not preclude other mechanisms such as decay also occurring; as we discuss further below. Forgetting, like other natural causes of loss such as erosion, can have many causes. For that reason, our aim in this paper has been not just to establish the coherence and plausibility of population dilution as a model of forgetting, but also to illustrate the different perspective this approach affords. A directly relevant example of how the PD model leads us to revise our perspective of specific issues this here is to be found in Wixted (2004, Figure 3), where Simon’s (1966) contrasting speculations as to Jost’s law – either higher levels of learning generate lower rates of forgetting, or lower levels of forgetting are related to the passage of time, or both – are considered. The PD model reinterprets these two explanations: Higher learning results in lower proportional forgetting because the memory population is larger and the constant dilution rate has proportionately less impact. Similarly, the passage of time reduces the rate of forgetting because, with increased delay, the population enlarges. In the context of the PD model, these two hypotheses are not contrasting, but different expressions of the same process. Overall, a mixed population, multiple trace model of forgetting affords different and interesting ways of looking at forgetting. In this final section we consider some of the broader issues arising from this theoretical approach and some of the difficulties of interpretation that accompany it.

*How do we interpret the model parameters?* In this article, the PD model has been applied to nine very different experiments to demonstrate the breadth of applicability
Dilution as a model of forgetting

of the approach. Here we consider what is to be learned from a comparison of parameter estimates across these experiments. This can be seen in terms of the additional insight we can gain from such comparisons. It is important to meet any concern that the PD model is overly complex or lacking in psychological relevance. The claim of being able to model recall, recognition and savings measures would be meaningless if the price of modelling so many experiments was an arbitrary use of free parameters and additional assumptions that applied to individual studies only that did not sum to a coherent view of forgetting. This is not straightforward because the design of any two experiments analysed in this paper differ on many dimensions; including: the testing methodology; the nature of stimuli and how they are presented; the delay periods over which retention is observed; whether there were repeated stimulus presentations; and the frequency of repetition where it occurs. The application of the PD model to these experiments therefore means that the sets of free parameters estimated in each case differ. This may be on the grounds of parsimony to enable a stronger test of the PD model, or because the experiment requires a specific adaptation of the model, as in the case of Fajnsztejn-Pollack’s data. As a result, for example, in only three experiments (Postman & Riley, Runquist, and Rubin et al.’s recall experiment) are estimates of the error term $E$ estimated independently. In the others, $E$ is determined \textit{a priori}, assumed to be zero, or, in the case of recognition experiments, not required in the model.

A robust analysis of all the parameter estimates is therefore not possible, but there are two sets of parameters where comparison is worthwhile. The first of these is to compare the rate of forgetting that was estimated for all experiments. The second is to look at those experiments where different degrees of learning are manipulated by
Dilution as a model of forgetting
differing repetitions of stimuli; looking at the impact of this upon the parameter $C$ in Equation [1] that estimates the proportion of traces in the population that support correct recall. These parameters are collated across the experiments in Table 2\textsuperscript{16}.

Insert Table 2 about here

\textit{i) Estimates of forgetting rates.} In the PD and PDE models, the parameter $k$ estimates, relative to the encoding of the original stimulus, how many null traces are added to the trace population as a function of time and is the principal determinant of the rate of decline of the retention function in the PD and PDE models. In its most developed expression – $2P_n\Theta/P_{sp}$ – it is a complex entity and comprises three elements. The first two of these can effectively be treated as simple constants: the Weber fraction $\Theta$ determines the length of the temporal record accessed in retrieval as a function of time, quantified as $2\Theta T$, where $T$ is the time elapsed between stimulus presentation and testing. The ratio $f/sp$ represents the density of irrelevant traces accessed per unit time within that record, that has been expressed throughout this paper as a fraction of the stimulus traces generated at stimulus presentation for the least overlearned condition in a given experiment (for which $sp$ – otherwise referred to as $C$ - is fixed at a value of 1). We note here that this ratio is critically dependent upon the value of the denominator $sp$. In the extreme, if stimulus encoding is very sparse, values of $k$ will therefore be driven high. Therefore, rates of forgetting will be dependent not only upon the density $f$ of interfering items between presentation and retrieval, but also upon those experimental circumstances that determine the strength of encoding. Finally, the ratio $P_n/P_s$ represents the notion that the population of null
Dilution as a model of forgetting

and interfering traces accessed by retrieval cues differ in their distribution of strengths in comparison to stimulus traces and that $P_n$ and $P_s$ represent the proportion of null and stimulus traces that exceed a criterion strength $\lambda$. In all recall experiments, and in Shepard and Strong’s recognition experiments, this theoretical elaboration is unnecessary and the ratio $P_n/P_s$ is effectively treated as unity. In the Fajnsztejn-Pollack and Rubin et al. recognition experiments, where varying false alarm rates make it necessary, the parameters $P_n$ and $P_s$ are estimated independently of $k$. In Table 2 we have crudely evaluated $\Theta$ as 0.05, as before, and normalised all observed values of $k$ to a common time unit (days) to express the observed values as multiples of $f/sp$ per unit time as the basis of comparison between experiments.

Despite the complexity of $k$, these estimates are suggestive of a variation of forgetting rates between experiments. Indeed, with the highest rates being over 200000 times greater than the lowest, some may find these variations implausibly wide; the most striking comparisons being between the high estimates of $k$ seen in Postman & Riley and in Rubin et al.’s experiments, and the low estimates seen in Shepard and Fajnsztejn-Pollack’s picture recognition experiments. More problematically, the values of $k$ illustrated in Table 2 indicate that those experiments conducted over short intervals generally show the highest values of $k$ and those conducted over the longest intervals the lowest. On the face of it, this is consistent with a view that the effective rate of forgetting declines with time faster, and in a different way, to that described by the PD model. There are therefore two questions to consider: Does the rate of forgetting vary between experiments? and, Does the comparison of $k$ across experiments point to a fatal weakness in the PD model?
Dilution as a model of forgetting

We offer two arguments to suggest this pattern is not problematic for the PD model and that forgetting rates probably do differ. Our first line of argument is to consider whether the variation in estimates of \( k \) are predictable from the experimental methodologies alone. Take, for example, the comparison of Rubin et al.’s recognition experiment with that of Fajnsztein-Pollack. In Fajnsztein-Pollack’s experiment, two separate sessions of over 30 minutes each were required on consecutive days simply to present the stimuli (at about 5 seconds per stimulus with 15 seconds between repeated stimuli), and recognition is tested over the following 49 weeks. By contrast, in Rubin et al.’s recognition experiment, the stimulus presentation and testing at all delays was completed in a continuous procedure taking about 47 minutes on a single day. This encompasses 630 trials at 4.5 seconds per trial, including feedback and intertrial intervals. As a result, the absolute density of records in memory as a function of time, including opportunities for extra-experimental interference, differs widely between these two experiments. In the Rubin et al. experiment, the participants can be regarded as in a steady state in which the memory record is exclusively filled with relevant events, whereas in Fajnsztein-Pollack’s experiment, many days and weeks go by during the course of the experiment in which the participant is carrying out tasks and experiencing stimuli quite unrelated to the experiment. Coupled to this, there are marked differences in stimulus materials. In Rubin et al.’s recognition experiment the stimuli comprised of just nine digits and five consonants in digit-letter-digit trigrams. The memory record over the entire retention period tested is therefore saturated with highly confusable stimuli. In comparison, Shepard and Fajnsztejn-Pollack’s picture recognition experiments used pictures chosen to be distinctive and of low confusability and the appearance of potentially
interfering events in the memory record will be relatively rare. In crude terms, the comparison of forgetting rates in our two examples is therefore like comparing the use of high and low frequency words such as ‘the’ and ‘theodolite’ as stimuli in different experiments and then considering potential sources of extra-experimental interference to these words from common language usage. If the stimuli are of high frequency and presented in a continuous block, extra-experimental interference will be commensurately high and it follows that forgetting will be faster. Conversely, stimuli which are distinctive, and for which potentially interfering experiences are rare in the delay between presentation and testing, will show slower rates of forgetting per unit time. Hence, one might argue, Rubin et al. selected their stimuli and methods precisely to facilitate a considerable amount of forgetting in the short period of their experiment, whereas Shepard and Fajnsztejn-Pollack selected their stimuli to deliver similar degrees of forgetting over the days and weeks of their experiment.

We also note that normalising $k$ to a common time unit of one day, although useful for purposes of comparison, is potentially misleading in the case of those experiments conducted over shorter periods. Indeed, reinterpreting what estimated values of $k$ actually mean in context is revealing. For example, Rubin et al.’s recognition experiment took less than an hour from start to finish, but the normalised value of $k$ represented in Table 2 effectively quantifies the interfering traces introduced as if this procedure had been run continuously over a matter of days. If we recalibrate $k$ to ask how many interfering traces are introduced in a given delay during the course of the experiment, the results are entirely plausible within the account of extra-experimental interference we have given. For example, in the 450 seconds it takes between presentation of a stimulus and its test at 99 trials later (the
Dilution as a model of forgetting

longest delay Rubin et al. used), the memory record accessed at retrieval is assumed to be about 45 seconds in duration. Scaling the estimate of $k$ in this experiment to this period reveals that approximately 6 interfering traces have been introduced into the population of retrieval traces. This is very close to the number of stimuli actually presented to the participant in the same time. In an experiment in which participants are completely engaged in processing stimuli and trials, this number is consistent with the view that stimuli and trials are a source of interference, and that that interference is the principal cause of forgetting. In these terms, there is no reason to think the high value of $k$ in this experiment and its comparison to lower estimates is of any particular theoretical consequence in itself.18

Our second line of argument concerns whether a more direct test of the stability of $k$ as a function of time can be established. More specifically, does $k$ always reduce with age of memories? Demonstrating that it does would suggest that the PD model had failed to capture an important element of forgetting. Identifying two experiments of equivalent $k$ but different time units (for example, hours compared to weeks) might provide such an argument if they could be found, but beyond the experiments provided here, this has not been possible. Such comparisons may also be of dubious value when, as we discuss below, experiments differing on timescales also allow additional processes such as rehearsal to render the comparison more difficult. However, even if such a comparison had been possible, the argument is not as convincing as it might initially appear. Demonstrating that two different experiments share a similar value of $k$ is not compelling when Table 2 demonstrates that a number of experiments with comparable timescales (Krueger, Slamecka & McElree, Runquist, Shepard, Strong) show different estimates of $k$ and we have already
Dilution as a model of forgetting

indicated above why experiments should be expected to vary as a result of their stimuli and the likely incidence of interfering experiences. The similarity of \( k \) between any two experiments of differing time course is therefore not strong evidence in itself of the stability of \( k \) over time.

What is possible is to return to the experiments analysed, where the data permit, to ask what happens to the estimates of \( k \) if they are estimated independently for different delays in a given experiment. If the argument that \( k \) declines with age holds – and with it a substantial issue about the validity of the PD model – this should be evident in all experiments. In practice, this test can be usefully conducted in four of the experiments analysed in this paper: Krueger, Postman & Riley, Rubin et al.’s recall experiment and Fajnsztejn-Pollack’s recognition experiment. In these experiments in the relaxation of constraint is least likely to produce over-prescription of the data and with it, unreliable estimates of \( k \) as a function of time.

The outcome of this analysis is complex, but does not support the view that variations in \( k \) show a consistent decline with delay. Specifically, Krueger’s data indicate a substantial increase of \( k \) with time, Postman & Riley and Rubin et al.’s experiments show no clear trend, whereas Fajnsztejn-Pollack’s recognition experiment does show a strong decline of \( k \) with delay. Only in Rubin et al.’s experiment it is possible to test the reliability of these trends. Here, the decline in a goodness-of-fit statistic for the 6 free parameters introduced by allowing \( k \) to vary for each of the seven delays modelled is not significant, \( \chi^2 (6) = 11.60, p > .05 \). In sum,
there is little support for the view that $k$ declines as a simple function of time in these experiments.

Assuming they represent significant trends, how best do we account for these divergent patterns in the relationship between forgetting rates and delay? As we have stated previously, we do not deny the possibility of many mechanisms contributing to forgetting alongside those described by the PD model, and a number suggest themselves. For example, if participants rehearse stimuli, as described in Equation 2 above, the effect of this would be to decrease $k$ with time, since its effect is to increase the representation of correct traces in the memory record. Conversely, if interference were to increase with time or stimuli were subject to decay processes, then $k$ would be expected to increase with delay as the PD model fails to predict the correct proportion of correct traces in the sampled memory population. It is noteworthy, therefore, that those experiments (Postman & Riley and Rubin et al.’s recall experiment) that prevent rehearsal, hold interference constant (because the participant is in effect in a steady state of high interference), and are conducted over relatively short duration, show stable estimates of $k$ with time. Beyond that, we can only speculate about the additional processes that have influenced the estimates of $k$ as a function of delay in these four experiments. Nevertheless, it is reasonably clear that they do not support the view that forgetting rates decline systematically with time in a way that fundamentally undermines the plausibility of population dilution as a mechanism for forgetting.

To summarise this analysis of the forgetting parameter, we are confident that the observed variation in forgetting rates we have demonstrated across experiments is
Dilution as a model of forgetting

entirely plausible given the different experimental circumstances to which the PD model has been applied. The model we propose for the retention function is an interference model that depends principally upon the density of irrelevant traces accessible in the memory record relative to the density of stimulus representations for its prediction of the forgetting function. In the explanatory terms of our model, this density is mostly determined by procedural, not biological factors. These include the type of stimuli and the subsequent exposure of the participant to similar material. The estimates of $k$ in Table 2 vary predictably in relation to the experimental designs and stimuli. To sustain the alternative hypothesis that all forgetting rates are equal, we would need to assume, amongst other things, that the density of stimulus traces in the memory record varies in such a way as to exactly counteract these experimental variables, and there is no reason to make that assumption.

\[ ii) \text{ Modelling the effects of overlearning.} \]  

In the experiments by Krueger, Postman and Riley, Fajnsztejn-Pollack, and Runquist, retention functions are compared between conditions in which stimuli were repeated a different number of times. For each experiment, the PD model estimates $C_n$ (the number of correct memory traces emanating from the $n$ repetitions of the stimulus) relative to the condition with the fewest number of learning trials, for which $C_n$ is fixed at 1. These estimates are also illustrated in Table 2. As one would expect, this illustrates that repeated presentations of the stimuli are always reflected in the PD model as an increase the estimate of the number of correct traces in the memory population (i.e., $C_n > 1$ for all $n > 1$). Where two levels of overlearning are present in the same experiment, a greater number of stimulus repetitions always produces a higher value of $C_n^{19}$. Any other pattern would have undermined the coherence and credibility of
the PD model, but it is important to note that the pattern of observed estimates is psychologically plausible, and as expected.

The other reliable characteristic of these parameter estimates is that $n$ repeated presentations produce values of $C_n$ disproportionally higher than $n$. Thus, for example, in Fajnsztejn-Pollack’s experiment, two presentations of a picture produces an estimate of 5.88 traces for every one produced with a single presentation. This finding is not new. For example, in Melton’s (1970) description of the spacing effect, the asymptotic level of recall for two presentations (that is, when spaced sufficiently widely apart to avoid the lowering effect of closely spaced repetitions) is invariably higher than that predicted from the independent function of two single traces. Under specific conditions, repetition enhances recall more than might be expected from two presentations of the stimulus.

A number of explanations for this finding are possible. One is that early learning trials are inefficient compared to later trials. If so and ignored, fixing $C_1 = 1$ inflates the estimate of early learning and drives up values of $C_n$ in the optimisation process as a result. Precisely this argument is seen in our analysis of the Postman and Riley experiment, where no disproportionality between $C_n$ and $n$ is seen provided the first $m$ trials are assumed to be ineffective. Nevertheless, the values for $C_n$ in Table 2 are seen as disproportionate when normalised to $C_1 = 1$. An alternative explanation is to suppose that repeated presentation not only increments the number of stimulus presentations, but also allows the retrieval of the previous instances of the same stimulus and replicates them in the temporal record. If recalls act as rehearsals that increment the trace population, it is easy to see that the trace population will expand.
Dilution as a model of forgetting

faster than the number of repetitions. In the extreme, for example, if every stimulus repetition invariably accessed and re-recorded all previous instances, the trace population \( C_n \) at the \( n \)th repetition would grow according to the relationship: \( C_n = 2 \cdot P_{n-1} + 1 \); giving the series 1, 3, 7, 15, 31 … for the growth in \( C_n \) as a function of \( n = 1, 2, 3, 4, 5 \ldots \) respectively.

In fact, even this expansion does not exceed the disproportionality observed in Table 2, and two other points emerging from Runquist’s and Rubin et al.’s recall experiments also suggest that yet other issues apply. Looking at Runquist’s experiment, the effect of testing half the stimuli after a short delay is anomalous in two respects. First, the estimate of \( C_{\text{test}} \) (modelling the impact upon the trace population of recalling a stimulus correctly on retest) is much greater than \( C_3 \); the effect of 3 repetitions of the stimulus in the initial stimulus presentation phase (9.55 compared to 2.12). \( C_{\text{test}} \) is also estimated at the same value whether the stimulus was presented once or three times. The memory test has a different effect upon memory compared to a repetition of the stimulus. Second, given that \( C_{\text{test}} = E_{\text{test}} \) (that is to say, a failure to recall correctly adds as many error traces to the memory population as does a correct recall add correct traces), we can rule out any explanation that successful retrieval has a privileged effect upon memory (unless we make the additional assumption that all errors made at retest are retrievals of errors learned at original presentation; which seems unlikely). The impact upon memory in this case therefore reflects the benefits of the temporal distribution of the added retest (e.g., see Landauer & Bjork, 1978), or an additional effect of the effort required in attempting recall; successful or otherwise. Either way, in Runquist’s data, the numerosity of repeated presentations is insufficient in itself to completely describe the effect upon
Dilution as a model of forgetting

the memory population of repeated presentations of the stimulus and interposed tests for recall.

Rubin et al.’s recall experiment offers a different challenge to the PD model. Here, consistent with the other experiments, higher levels of performance are associated with a systematic and disproportionate rise in the estimate of $C_n$. The difficulty of interpretation here is that the difference in performance of the five groups does not result from an experimental manipulation of the number of stimulus presentations, but is defined *post hoc* by dividing the participants into five groups according to their overall memory performance in the experiment. Variations in $C_n$ must therefore be explained in terms of differences between participants rather than between experimental conditions. Such differences could arise from a number of factors, all of which emphasise different aspect of encoding efficiency such as: motivation; speed of processing; expertise; and the degree to which participants engage in elaboration as opposed to maintenance of stimuli (Craik & Watkins, 1973).

Taking our analysis of Runquist and Rubin et al.’s experiments as support for the PD model therefore requires an element of special pleading. The general problem for the PD model is how to relate predictions based upon the numerosity of traces to the enormous literature relating memory performance to the close dependency of encoding strategy, what is encoded, and retrieval mechanisms (e.g., Morris, Bransford & Franks, 1977). For example, how are we to describe what happens when Runquist’s participants make their first attempt to recall stimuli shortly after presentation? And, what are Rubin et al.’s highest performing quintile group doing differently from the lowest? Is it psychologically plausible to treat these processes purely in terms of the
number of traces produced, or are important aspects of the retrieval process are being disregarded thereby that will change our view of forgetting? The theoretical relationship between the nature of encoding as a function of experimental circumstance, and the numerosity of traces as a function of repetition remains a central issue in the study of memory, to which our contribution would seem to be that numerosity is an extremely useful theoretical construct for modelling the forgetting function.

Revisiting some assumptions of the PD model. Arguably, the impact of some of the assumptions we have made to present the PD Model cannot be felt in model fits to individual experiments, or in the parameter estimates directly. In this section we consider four of these: the validity of concentrating upon long term forgetting only; the exclusion of decay and consolidation process in the PD model; the assumption that retrieval does not allow resampling (i.e., it is a one-shot process); and finally whether alternative models of dilution can be considered.

i) One function or two – is it valid to exclude short term memory? Many experiments collect forgetting data over a range of delays from the immediate to many weeks. Nor is it unusual for forgetting functions to be represented graphically in log-log form with straight lines representing reasonable approximations of the retention curve over the entire time span (e.g., Begg & Wickelgren, 1974; Rubin et al., 1999; Sikström, 2002), encompassing very short as well as long retention intervals. These linear relationships imply the application of a power law model to the retention function, and this law has the attraction of approximating findings across a wide array of learning tasks (see Anderson, 2000, Chapter 7). They are therefore attractive on
Dilution as a model of forgetting

grounds of parsimony. Further, although (as noted above) they are problematic in the face of the reliable systematic deviations observed by Rubin et al. (1999) and others, it happens that incorporating the entire span of delay from very short to very long delays (as opposed to excluding short delays as we have done here) can superficially improve the fit of a linear function to the entire data if higher levels of short term recall gravitate toward the linear function. The analysis of Strong’s (1913) data above, the only analysis where we depart from our policy of excluding short term memory, illustrates this in Figure 10.

Nevertheless, we argue that fitting linear retention functions across all delays in log-log transformations is potentially misleading on two counts. First, it effectively implies a unitary model of forgetting across all delays, which is contentious. Second, even if this were acceptable, the empirical property in which most, if not all, fits of the power law underestimate recall at intermediate delays and overestimate the same at longer delays still demands explanation, even if the sigmoidal deviation from the linear is perceived as small. The PD model provides such an explanation, as do, from different perspectives, Rubin et al. (1999) and Sikström (2002). As it happens, the validity of this assumption is not, in itself, central to the acceptability of our model. Since the aim of this paper is to make a prima facie case for dilution as a model for long term forgetting, the issue of whether or not short term memory data are excluded is not one upon which the idea stands or falls, although it is obviously interesting to speculate how the principles described here might apply to the analysis of short term recall.
The alternative strategy, evident in the details of Rubin et al.’s (1999) sum of exponential model and possibly in Sikström’s (2002) connectionist model, is to add additional assumptions modelling short term recall to enable a fit of the retention function to data from all delays. As a rhetorical device, we adopted precisely this approach in the analysis of Strong (1913) to make contact with earlier analyses of Strong’s data and to exemplify some of the issues in so doing. However, our preference here has been to preserve the simplicity of our basic model and indeed the purity of our tests of it, by choosing not to model short term data. Do we need to defend this choice? The answer would be yes, if one believed that replacing the parsimony of the power law with two processes – one for the short term and one for the long is unjustified. But the parsimony of the PD model in relation to its competitors belies this, and the evidence for an operational distinction between processes of short and long term memory seems sufficiently compelling (e.g., see Roediger, Marsh & Lee, 2002 for a review) to allow us to partition and analyse only long term data in this way.

ii) Decay, consolidation, and the PD model. Our simple formulation of the PD model relates retrieval entirely to the statistics of a population of all-or-none traces. This population is defined by the initial conditions of learning and the time elapsed before testing. The PDE model further adds some account of the distribution of strength in a population of traces, but neither model incorporates directly the further possibility that individual traces (be it C, E, or null) either consolidate or decay, despite evidence in the neuroscience literature that these processes may be important in the dynamics of neural function. How and why can the PD model ignore these processes?
Dilution as a model of forgetting

A simple answer is that it has not been necessary to make any further assumptions about the fate of individual traces in populations to predict the observed retention functions with the PD model. Also, by largely ignoring short term memory processes, we may have made our task easier by finessing the issue as to how and what traces consolidate in the transition from short to long term representations. More importantly still, provided we remain restricted to long term retrieval, issues of consolidation and decay are not directly relevant to the modelling of the retention function within the framework of the PD model. If we assume that $C$, $E$ or null traces are equally prone or resistant to decay as a function of time (which is reasonable given that they are of comparable age and are categorised by experimental definition), the PD model is indifferent to decay because in the retention function it is predicted on the basis of the relative proportion of correct traces within a trace population. That is, provided the elements within Equation [1] decay at the same rate, it is easy to see that the predictions it makes are unchanged by decay. The PD model does not deny the possibility of trace decay as a forgetting process, but as a first approximation, its effects are not necessarily detectable in the retention function.

That said, we can identify three ways in which decay could still have a detectable effect upon recall that could be exploited in future research. First, any decline in the size of the trace populations could, at the limit, have an impact upon the variability of the probability of correct retrieval with delay if the frequencies of $C$, $E$ or null traces become so small that quantal effects are detectable. Although the technical difficulties in identifying the effects of this may be considerable, it may be
Dilution as a model of forgetting

possible to detect trace decay at extreme delays in its effect upon the variance in the data.

Second, it is an obvious and necessary step to extend the PD model to latencies of recall. If no decay is observed, the population size $P$ from which recall is sampled increases linearly according to the function $P = C + E + kt$. Although modelling of recall latencies has received less attention than has the forgetting function (e.g., see Ratcliff & McKoon, 2000), most models predict an increase in recall latency with increasing set sizes (e.g., Anderson, 1974; Wixted & Rohrer, 1993). However, to take a simple analysis, if these traces are subject to all-or-none decay with a probability of any trace being lost in unit time as $d$, then the population size as a function of test delay becomes the more complex quantity: $P = (1-d)(C + E + kt)$. This function initially increases and then declines with $P$ reaching a maximum at $t_{max} = 1/\ln(1-d) - (C+E/k)$. Depending upon the values of $d$, $C$, $E$ and $k$, $t_{max}$ is either negative (in which case the population only declines in real time), or positive. Therefore, if decay is occurring, we can speculate that some forgetting experiments will show a peak in recall latencies as a function of delay. In fact, such a peak is suggested in Rubin et al. (1999, Figure 6) in the analysis of recall latencies as a function of delay. We speculate that since $C$, $E$ and $k$ are potentially manipulable, it may be possible to conduct experiments in which predictions for $t_{max}$ can be tested and the decay rate $d$ estimated. Further analysis of recall latencies as a function of delay in forgetting experiments might reveal more of the dynamics of trace decay where the time course of the probability of correct recall cannot.

The third way in which decay may be having an indirect impact upon the retention function is seen in the PDE model in the additional process by which the
Dilution as a model of forgetting

memory population is reduced only to those traces above a criterion strength; with that criterion declining as a function of time. This process was added to the PDE model because of the commonly observed rise in false alarm rates with delay, and the general acceptance that criterion levels can shift adaptively (e.g., Singer & Wixted, 2006). An interesting question here is to speculate what this process means in terms of the PD model. Some explanations are meta-mnemonic, and tangential to the PD model. One possibility is that criterion shifts reflect the participant’s confidence in recall as performance declines. Another is that, with recognition rates declining at longer delays, participants lower their threshold to enable more ‘old’ responses to balance the otherwise high proportion of ‘new’ responses they will make if recognition fails. But this process could also reflect an adaptive response to decay. In the initial stages of the experiment, a high criterion selects only strong candidates for recognition. Thus in the Fajnsztejn-Pollack data, a criterion value of 4.23 at one week effectively rules out all distracters; whose strength is represented by the standardised distribution $N(0, 1)$. After 49 weeks, the criterion of 0.89 effectively means that most of the target strength distribution $N(m, \sigma)$ has been included as well as a significant proportion of distracters. But in the PD model framework, there is no adaptive logic to increasing the overall trace population in this way if the effect of this is to decrease the likelihood of recall by further diluting the trace population. The logic would surely point to retaining high criteria levels to exclude as many irrelevant traces as possible in the memory population. If that is correct, the need to model an additional decline in strength criterion in the PDE model suggests that the target population itself is declining in strength and the criterion shift represents a response designed to maximise retrieval. A closer investigation of the relationship between delay and
observed criteria shifts may offer further insights on how processes of decay may operate within the PDE model.

In the PD model, as with decay, there is no need to model consolidation. The characteristics of the retention function it has been used to explain (e.g., Wixted, 2004) are otherwise explained, where necessary, by proposing a multiplicity of traces. Thus, overlearning ‘consolidates’ memory (by which we mean operationally that it increases the likelihood of a correct response at some time in the future) simply by generating more correct traces, thereby increasing their proportion in the overall trace population. Consolidation can also be seen as the corollary of decay – those traces that consolidate are less prone to decay. At first glance, consolidation would seem to be more detectable than decay because it would result in retention functions with asymptotes significantly above zero. However, establishing that non-zero asymptotes represent consolidation is difficult. First, Wixted (2004) demonstrates that even in Bahrick (1984) – data held as exemplifying a consolidated ‘permastore’ – it is not safe to assume that retention is not progressing to a zero asymptote; even if that is hundreds of years in the future. Second, in data such as Runquist (1983, Experiment 2), proving that non-zero asymptotes represent anything more than guessing is difficult. Sophisticated guessing strategies in cued-recall experiments mean that participants need remember very little information to be able to guess correctly at levels significantly higher than the supposed chance level (Lansdale & Laming, 1995) – levels that would also raise asymptotic values of the retention function above zero. One might suppose that this latter technical problem could be overcome by some form of repeated-test R-T-T ... design, because consolidated traces would be expected to produce recall reliably in such situations, whereas guesses would be intrinsically
variable. However, Lansdale and Laming also demonstrate that repeated tests of memory serve to generate memories in their own right, confounding a test of consolidated traces with recall for previous trials. To summarise, the PD model does not rule out the possibility of decay, or that traces differ in their vulnerability to decay. But, in the retention function at least, it claims that their effects are empirically negligible and technically difficult to evaluate.

**iii) One-shot retrieval.** Except in the case of the 2AFC recognition experiment, the PD and PDE models assume that retrieval in forgetting experiments is a one-shot process. That is, the outcome is determined by the first attempt only, and resampling of memory is not permitted. For this reason amongst others, in this article we have not considered free recall experiments; where this assumption is clearly violated. In cued recall particularly, although the blocking effects of initial recalls is a well-known phenomenon, it seems counterintuitive to suppose that if recall fails, memory is not resampled. How reasonable is this assumption and is it essential to the PD model?

We do not find this assumption implausible for the kinds of experiments analysed here. Numerous examples in the literature point to the inhibiting effect of prior recall on later attempts (e.g., Brown, 1968). This is referred to by Reason and Lucas (1984) as the ‘ugly sister’ phenomenon. Resampling is a less efficient process than the initial attempt at retrieval. Furthermore, while in some experiments participants can be expected to try hard to resample memory, there is no reason to think that this applies to experiments on forgetting, where the participants themselves will have the expectation that some information will be lost and may regard prolonged attempts at recall as pointless. Finally, in many of the experiments, the number and
Dilution as a model of forgetting

pacing of the trials are such as to question whether the participants would have had much time to resample memory. Overall, given the expectation of failure and the cognitive cost of attempting to resample on failure with low expectations of a successful retrieval, we do not find it at all implausible to suppose that, in these experiments at least, recognition and cued recall can be modelled as one-shot processes. That notwithstanding, an interesting test of this would be to compare experiments in which resampling was or was not encouraged, to examine the impact of these processes upon the retention function. Such a comparison would be preliminary to the more complex issue of how the PD model would be applied to free recall, where the cues for recall are covert and the participant is more likely to resample memory.

It is revealing to examine the effects upon the PD model if the assumption of one-shot recall were violated. Suppose, as an alternative, we assumed in Equation [1] that whilst retrieval of a C or E trace terminated recall (E as well as C because the participant has no way of discriminating which is correct), null retrievals resulted in a resampling. In the extreme, if this resampling were to continue until an E or C trace was sampled, Equation [1] becomes \( p(R) = C/(C + E) \). If a hybrid model is assumed in which some resampling is permitted (because otherwise forgetting would not be a function of time), it is easy to demonstrate that the effect must be to lower the estimate of C compared to those presented in this paper (when only one attempt at sampling is allowed). It is therefore conceivable that the disproportionality observed in our estimates of \( C_n \) derive from a violation of this one-shot assumption. As a simple test of this possibility, we re-fit Krueger’s data illustrated in Figure 3 with a hybrid PD model. In this hybrid model, in which participants resample memory with
Dilution as a model of forgetting

probability \( x \) until a \( C \) or \( E \) trace is accessed and with probability \((1-x)\) perform according to Equation [1] as before, the best fit occurs when \( x = 0 \). This leaves the original fit of Krueger’s data unchanged. Also, if \( x \) is fixed at a small but finite value (which results in a poorer fit to the data), the impact is felt not upon the values of \( C_n \), which barely change, but upon \( k \). We therefore conclude that the assumption that memory is sampled once is a reasonable one to make, in these studies at least, and that a violation of that assumption does not otherwise explain the observed disproportionality in values of \( C_n \) as a function of repeated presentations of the stimuli.

\( iv \) Alternative models of dilution. In the PD model, we choose to represent the process of dilution – effectively a leakage of null traces into memory populations modelled by the addition of \( D/C \) per unit time \( t \) – as a process determined by the temporal organisation of memory rather than leaving it as an atheoretical mathematical abstraction. In this section we explore the question: Are there theoretical variants that have the same empirical consequences?

It is impossible to say definitively that alternative models will not prove to be mathematically equivalent to the PD model in the future. However, some simple alternatives can be excluded. For example, suppose the gradual diminution of the effect of \( C \) and \( E \) traces in the population, and the increase in null traces, was explained by a transition process in which traces decayed to leave accessible 'husks' with no content. This decay-of-content model leaves the overall trace population of \( C + E \) unchanged, but sees the likelihood of accessing \( C \) traces surviving to time \( T \) declining according to a function of the form \( p(R) = C(1-k)^T/(C+E) \). This is an
Dilution as a model of forgetting

exponential function consistently rejected in previous studies and it would not be expected to be adequate here. As a matter of comparison with the PD model, applied to Rubin et al.'s (1999) data this model, again using four free parameters, provides a goodness-of-fit statistic of $\chi^2(35) = 133.3$; a much poorer fit than that of the PD model.

Another mechanism we can probably exclude as a plausible explanation of dilution is the concept of 'contextual drift' (Gillund & Shiffrin, 1984). In this formulation of the SAM model\(^{21}\), forgetting as a function of time is explained as an increasing drift from the learning context with an increasing detriment in recall. This seems a particularly attractive candidate mechanism for dilution. In terms of the PD model it seeks to explain the introduction of null traces by suggesting the new context cues associated with drift allow access to irrelevant traces at recall. This idea has resonance with other approaches to recall, but it is unlikely that this process is mathematically or psychologically plausibly equivalent to the PD model. For one thing, it would require context drift to have an unbounded range of potential drift to explain the forgetting function across the range of possible delays. This seems implausible. For another, it would need to explain why context drift introduced irrelevant traces, but left the accessibility of $C$ and $E$ traces unchanged. Finally, it would need to explain why context drift predicted a linear influx of null traces as a function of time whilst most sampling models would suggest other functions such as the exponential (e.g., Estes, 1955).

**Concluding Remarks.** Our aim in this paper has been to establish the distinctiveness and potential of the PD model as a theoretical contribution (adding diversity to the
Dilution as a model of forgetting

collective theoretical gene pool, so to speak), rather than establishing its superiority
over rival approaches. In any case, the nature of forgetting data makes the latter aim
extremely difficult in practice (see Rubin & Wenzel, 1996; Rubin et al., 1999). We
have taken the view that the complexity of trying would obscure our main points with
the minuta of necessarily complex and near-intractable analyses (see Brown et al. for
similar comments, or Laming, 1992, for a different approach to this problem of
exposition). This has motivated our strategy of presenting simple (but adequate)
models for cued recall before the more complex models required for recognition
paradigms. For similar reasons, we have concentrated upon forgetting in the long
term (see also Wickelgren, 1972); excluding short term processes in our analyses. We
have also argued that, while reasonable closeness of fit is a prerequisite to taking a
model seriously, it is not, in itself, sufficient to be persuasive (see Roberts & Pashler,
2000). We hope that the simplicity of the basic processes in the PD model, the
characteristics of forgetting it successfully accounts for, and the research agenda it
implies, serves that purpose.
Dilution as a model of forgetting

References


Dilution as a model of forgetting


*Journal of Verbal Learning and Verbal Behavior, 12*, 599-607.


*Psychological Review, 93*, 411-428.
Dilution as a model of forgetting


Dilution as a model of forgetting


Dilution as a model of forgetting


Dilution as a model of forgetting


Dilution as a model of forgetting


Dilution as a model of forgetting


Dilution as a model of forgetting

Author Note

Mark W. Lansdale, School of Psychology, University of Leicester, and Thom Baguley, Division of Psychology, Social Sciences, Nottingham Trent University, United Kingdom.

The authors would like to thank David Rubin, Sean Hinton, and Amy Wenzel for making their data available for reanalysis and also Anthony Sackfield for clarifying the proofs presented in Appendix A of this article. We are also grateful to David Rubin, John Wixted, Geoffrey Loftus and Sverker Sikström for their detailed and constructive comments on earlier versions of this article.

Working implementations of all models reported here are available, in Excel™ form, at http://hdl.handle.net/2381/3850 and also at http://www.le.ac.uk/pc/ml195/.

Correspondence concerning this article should be addressed to Mark W. Lansdale, School of Psychology, University of Leicester, Leicester, LE1 9HN, United Kingdom.

E-mail: ml195@le.ac.uk
Dilution as a model of forgetting

Figure and Table Captions

Figure 1. An illustration of Jost’s second law (1897). Stimulus A is overlearned with respect to Stimulus B and learned earlier. Consequently, the probability of recalling Stimulus B, declines faster than that of A and therefore the two functions intersect at delay $t$. In this diagram the dashed lines in the early stages of the retention functions for stimuli A and B reflect our concern in this article to model only long term retention functions.

Figure 2. The principal elements of the Population Dilution Model. Filled circles represent traces that will generate correct retrieval. Open circles represent traces that will produce an erroneous recall or null traces. The memory record shows a varying density of correct and error/null traces, with a particular concentration of correct traces at the time of stimulus presentations. A sector of this record, centred on the stimulus and of width $2\Theta T$ is selected and subject to filtering in relation to cue strength to determine a sub-population of traces which is sampled at random for recall.

Figure 3. A fit of the Population Dilution model to Krueger’s (1929) Experiment plotting ln(percentage recall) as a function of ln(days delay). Data points indicate observed recall for groups with 0%, 50% and 100% overlearning trials and the continuous lines represent the model predictions in each case.
Dilution as a model of forgetting

Figure 4. A fit of the Population Dilution model to Postman and Riley’s (1959) experiment plotting ln(number of items recalled) as a function of ln(number of intervening trials). Data points indicate observed recall for groups with 10, 20 and 40 learning trials and the continuous lines represent the model predictions.

Figure 5. A fit of the Population Dilution model to Slamecka and McElree’s (1983) Experiment 4 plotting ln(number of items recalled) as a function of ln(delay) in minutes. Data points indicate observed recall for groups in ‘high’ and ‘low’ learning conditions and the continuous lines represent the model predictions.

Figure 6. A fit of the Population Dilution model to Runquist’s (1983) Experiment 2 plotting ln(mean percent correct) as a function of ln(delay) in hours. Data points indicate observed recall for groups with 1 or 3 presentations of the stimuli and the continuous lines represent the model predictions. Model predictions for groups TC1 and TC3, being indistinguishable, are plotted as a single function. Note that the presentation of ln(percent correct) spatially exaggerates the variance in low-performing conditions.

Figure 7. Rubin, Hinton and Wenzel (Figure 4), showing probability of recall against lag in log-log form for 300 participants divided into quintiles, and the predictions for the Population Dilution model (continuous curves). Note that data for lag 0 are not shown and these and the next two shortest lag data to the left of the vertical dashed line are deemed to be influenced by short term recall and therefore excluded from modelling. Therefore the PD model is applied only to data from the longer lags shown to the right of the vertical dashed line.
Figure 8. The hypothetical distribution of trace strengths for foils (continuous line) and targets (dashed line) in the PDE model. Foil strength is assumed to be normally distributed with a mean strength of 0 and standard deviation 1. Target strength is the weighted sum of the foil distribution and a target distribution of mean $m$ and standard deviation $\sigma$. These distributions are not necessarily equally weighted (depending upon experimental circumstances) and here are illustrated in a ratio of 3 (target) to 1 (foil).

Figure 9. A fit of the PDE model to Shepard's (1967) picture recognition data fitting ln(proportion correctly recognized) as a function of ln(elapsed time) in days.

Figure 10. A fit of the Population Dilution model to Strong's (1913) recognition data plotting proportion of targets recognized as a function of ln(elapsed time) in minutes.

Figure 11. A fit of the PDE model to the Adolescent Data from Fajnsztejn-Pollack (1973). Note that whilst data for 2 or 4 repetitions was collected from one group (A or B) only, false alarm (FA) and percentage recognitions for single presentations of stimuli were recorded from both groups A and B.

Figure 12. A fit of the Population Dilution model to Rubin et al.'s (1999) recognition data plotting ln(probability recognized) as a function of ln(elapsed lag).
Dilution as a model of forgetting

Figure 13. The predictions of a PD model simulation of Recall (Black circles) and Savings measures (white circles) in Krueger's (1929) experiment (Figure 13a) compared to the observed data (Figure 13b).

Table 1. Hit rates and false Alarm rates at each delay from Fajnsztejn-Pollack (1973); adolescent participants only.

Table 2 Summary of parameter values estimated across the experiments analysed.

a) $k$ – the forgetting rate normalised to represent the ratio of stimulus traces added to the trace population per day relative to the number of correct traces emanating from a single presentation, $C$, fixed at 1. Estimates from experiments in which delay was measured other than in days have been adjusted accordingly.

b) The relationship between the number of presentations of a stimulus, $n$, with estimates of $C_n$, the number of correct traces assumed to result in the PD model. Learning weights (usually determined by the number of presentations of the stimulus) have been normalised relative to the lowest weight (typically a single presentation) in a given experiment being fixed at 1. Values for the lowest performing group in each experiments (where $n = 1$, $C_1 = 1$ by definition) are not shown.

Table 3. Values of $k$ estimated for each testing delay for four experiments. Figures in bold type represent the delay in each experiment in time units given in parentheses. Large values of $k$ have been rounded to the nearest integer value.
Figure 1

Dilution as a model of forgetting

Delay since stimulus A was learned

Probability of recall

Stimulus A

Stimulus B

T

t
Dilution as a model of forgetting

Figure 2

Temporal characteristics define the segment of the memory record.

FILTERING BY CUE

Retrieval cues filter the memory record to those reaching threshold strength.

RESPONSE SELECTION

Retrieval is a random selection.
Dilution as a model of forgetting

Figure 3

![Graph showing the relationship between natural logarithm of days delay and natural logarithm of percentage recall. The graph includes three lines representing different conditions: 0 (pred), 50 (pred), and 100 (pred). The graph also illustrates data points for 0 obs, 50 obs, and 100 Obs.]
Figure 4

Dilution as a model of forgetting
Dilution as a model of forgetting

Figure 5
Dilution as a model of forgetting

Figure 6

Retention interval (ln hours) vs. mean ln(percent correct) for different models and observed data points.
Figure 7
Dilution as a model of forgetting

Figure 8

[Diagram showing dilution as a model of forgetting with axes labeled 'trace strength', 'Foil', and 'Target'.]
Dilution as a model of forgetting

Figure 9

![Graph showing dilution as a model of forgetting](image-url)
Dilution as a model of forgetting

Figure 10

![Graph showing dilution as a model of forgetting](image-url)
Figure 11

Dilution as a model of forgetting

![Graph showing dilution as a model of forgetting with Ln delay (weeks) on the x-axis and Ln p(correct recognition) on the y-axis. The graph includes lines for Model 1, Model 2, Model 4, and Model FA, as well as symbols for Obs Group B4, Obs Group A2, Obs Group B1, Obs Group A1, Obs FA Group B, and Obs FA Group A.]
Dilution as a model of forgetting

Figure 12

[Graph showing a log-log plot with the x-axis labeled as ln (elapsed trials) and the y-axis labeled as ln (proportion correctly recognised). The plot shows a downward curve with data points.]
Dilution as a model of forgetting

Figure 13a

![Graph showing the proportion recalled/savings versus elapsed delay (days).]

Figure 13b

![Graph showing another set of data with similar axes as Figure 13a.]
Appendix A

The population dilution model and Jost’s law. Suppose we construct an experiment, following Krueger’s (1929) anticipation method and illustrated in Figure 1, in which Stimulus A is trained to a criterion of $C$ correct anticipations after which the participant undertakes a further $\Delta$ overlearning trials. In the course of this training, we also assume that erroneous anticipations give rise to $E$ error traces. Assuming a time constant $k$, the Population Dilution model predicts a probability of recall at time $t$ of

$$p(A/t) = \frac{C + \Delta}{C + \Delta + E + kt}$$  \hspace{1cm} [A1]$$

At a later time $T$, a second stimulus B is trained to the same criterion but does not receive the overlearning trials, giving a recall function as time elapses since $T$ of

$$p(B/t) = \frac{C}{C + E + k(t - T)}$$  \hspace{1cm} [A2]$$

In order to demonstrate Jost’s Law as illustrated in Figure 1, we assume by a suitable choice of $C$, $E$, and $\Delta$ that $p(B) > p(A)$ at $T$.

The instantaneous rates of forgetting for stimulus A and B at time $t$ are given by

$$\frac{dp(A)}{dt} = -k \frac{(C + \Delta)}{a^2}$$  \hspace{1cm} [A3]$$

$$\frac{dp(B)}{dt} = -k \frac{C}{b^2}$$  \hspace{1cm} [A4]$$
Dilution as a model of forgetting

Where \( a = C + E + \Delta + kt \) and \( b = C + E + k(t - T) \). The retention functions \( p(A/t) \) and \( p(B/t) \) intersect at point \( t_i \) where

\[
\frac{C + \Delta}{a} = \frac{C}{b} \tag{A5}
\]

The difference in instantaneous forgetting rates \( D = \frac{dp(A)}{dt} - \frac{dp(B)}{dt} \) and at \( t_i \) is

\[
D = k \frac{C}{b^2} - k \frac{(C + \Delta)}{a^2} \tag{A6}
\]

Since from \( \text{[A5]} \) \( a = b(C + \Delta) / C \), \( \text{[A6]} \) reduces to

\[
D = k \frac{C}{b^2} \left( 1 - \frac{C}{C + \Delta} \right) \tag{A7}
\]

Which is necessarily positive unless \( \Delta = 0 \); indicating that the younger trace necessarily decays more quickly than the older.

The population dilution model, the power law, and log-log plots. It convenient and common (e.g., Anderson, 2000) to present retention functions relating the probability of recall and delay transformed as natural logarithms. In this case hypothetical forgetting functions complying with a power law of the kind \( p(R) = xt^y \) become linear (see also Rubin, Hinton & Wenzel, 1999, Figure 4). In contrast, the PD model predicts a more complex function to which we argue data in all experiments more closely correspond. Equation \([A1]\) above can be rewritten as
Dilution as a model of forgetting

\[ p(A/T) = 1/(1 + a + bT) \]  

[A7]

where \( a = E/(C + \Delta) \) and \( b = k/(C + \Delta) \).

Taking logarithms and writing \( y = \ln(p(A/T)) \) and \( x = \ln(T) \) gives

\[ y = -\ln(1 + a + be^x) \]  

[A8]

that is non-linear in \( x \) and \( y \), with an accelerating slope of

\[ \frac{dy}{dx} = \frac{-be^x}{1 + a + be^x} \]  

[A9]

Since \( a, b, \) and \( e^x \) are always positive, this curve is always negative with its steepest gradient as \( T \to \infty \) of -1 and shallowest gradient of \(-k/(k+P)\) at \( T = 0 \) (where \( P \) is the size of the memory population immediately after the final learning trials). Note however, that this asymptotic value is rarely approached in practice with even the longest-term experiments terminating well before that point is reached (e.g., see Wixted, 2004, p. 875). Also, this hypothetical value should not be confused with exponents estimated for power functions assuming a linear function across all delays which will naturally be considerably less since it takes no account of the curvilinear nature of the data and approximates a linear function only over the delays actually measured.
Dilution as a model of forgetting

Table 1

<table>
<thead>
<tr>
<th>Group</th>
<th>No. of repetitions</th>
<th>Delay (weeks)</th>
<th>Hit rate</th>
<th>False Alarm rate</th>
<th>Hit rate</th>
<th>False Alarm rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>2  5  10  20  49</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A 1</td>
<td></td>
<td></td>
<td>.51</td>
<td>.33</td>
<td>.37</td>
<td>.37</td>
</tr>
<tr>
<td></td>
<td>False Alarm rate</td>
<td>.02</td>
<td>.06</td>
<td>.13</td>
<td>.15</td>
<td>.21</td>
</tr>
<tr>
<td>B 1</td>
<td></td>
<td></td>
<td>.44</td>
<td>.31</td>
<td>.31</td>
<td>.23</td>
</tr>
<tr>
<td></td>
<td>False Alarm rate</td>
<td>.03</td>
<td>.09</td>
<td>.08</td>
<td>.09</td>
<td>.17</td>
</tr>
<tr>
<td>A 2</td>
<td></td>
<td></td>
<td>.81</td>
<td>.71</td>
<td>.64</td>
<td>.49</td>
</tr>
<tr>
<td></td>
<td>False Alarm rate</td>
<td>.02</td>
<td>.06</td>
<td>.13</td>
<td>.15</td>
<td>.21</td>
</tr>
<tr>
<td>B 4</td>
<td></td>
<td></td>
<td>.51</td>
<td>.33</td>
<td>.37</td>
<td>.37</td>
</tr>
<tr>
<td></td>
<td>False Alarm rate</td>
<td>.03</td>
<td>.09</td>
<td>.08</td>
<td>.09</td>
<td>.20^</td>
</tr>
</tbody>
</table>

^ Note that this number does not correspond to the False Alarm rates in the B1 data; with which it should be common. This implies a clerical error either in reporting or in analysis. Further, attempts to re-compute values of d' from this table produces discrepancies with values reported in the original paper; possibly reflecting these clerical errors or rounding errors in calculation. The reported data must therefore be treated with this uncertainty in mind, but we judge it suitable for the present purposes.
Dilution as a model of forgetting

Table 2

a)

<table>
<thead>
<tr>
<th></th>
<th>k</th>
<th>stimuli</th>
<th>longest delay units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Krueger</td>
<td>46.95</td>
<td>12 word lists</td>
<td>days</td>
</tr>
<tr>
<td>Postman &amp; Riley</td>
<td>158710.44</td>
<td>CVC lists</td>
<td>trials</td>
</tr>
<tr>
<td>Slamecka &amp; McElree</td>
<td>12.36</td>
<td>16 sentences</td>
<td>days</td>
</tr>
<tr>
<td>Runquist</td>
<td>10.24</td>
<td>24 word pairs</td>
<td>days</td>
</tr>
<tr>
<td>Rubin et al.</td>
<td>50610.24</td>
<td>word pairs</td>
<td>trials</td>
</tr>
</tbody>
</table>

Recognition experiments:

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Shepard</td>
<td>0.79</td>
<td>612 pictures</td>
<td>days</td>
</tr>
<tr>
<td>Strong</td>
<td>5.44</td>
<td>20 word list</td>
<td>days</td>
</tr>
<tr>
<td>Fajnsztejn-Pollack</td>
<td>0.67</td>
<td>280 pictures</td>
<td>weeks</td>
</tr>
<tr>
<td>Rubin et al.</td>
<td>11668.20</td>
<td>DLD trigrams</td>
<td>trials</td>
</tr>
</tbody>
</table>

b)

<table>
<thead>
<tr>
<th>Study</th>
<th>n (normalised to $C_0 = 1$)</th>
<th>estimated $C_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Krueger</td>
<td>1.5</td>
<td>4.71</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>6.57</td>
</tr>
<tr>
<td>Postman and Riley</td>
<td>2</td>
<td>5.56</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>14.39</td>
</tr>
<tr>
<td>Fajnsztejn-Pollack</td>
<td>2</td>
<td>5.88</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>42.24</td>
</tr>
<tr>
<td>Runquist</td>
<td>3</td>
<td>2.12</td>
</tr>
<tr>
<td></td>
<td>Retest†</td>
<td>9.55</td>
</tr>
<tr>
<td>Rubin et al. (recall)</td>
<td>1.16</td>
<td>1.77</td>
</tr>
<tr>
<td></td>
<td>1.49</td>
<td>2.62</td>
</tr>
<tr>
<td></td>
<td>1.80</td>
<td>4.26</td>
</tr>
<tr>
<td></td>
<td>2.52</td>
<td>9.97</td>
</tr>
</tbody>
</table>

† In the text, we refer to this value as $C_{test}$, because it refers to the impact of a recall test (rather than an explicit re-presentation of the stimulus) on the population of correct traces. $C_{test}$ is the increment in the memory population of correct traces in the event of a correct recall, and $E_{test}$ increments the errors traces in the event of recall failure, the optimised values being the same in each case.

†† In the Rubin et al. recall experiment, which is included here for reasons of comparison, variations in learning weight are derived post hoc from overall levels of performance and not from manipulations of stimulus presentation. Thus n is calculated (relative to the lowest performing group) from overall levels of performance and $C_n$ estimates the number of correct traces required to support that level of performance.
Dilution as a model of forgetting

Table 3

Recall experiments

Krueger:

<table>
<thead>
<tr>
<th>Day</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>7</th>
<th>14</th>
<th>28</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimated $k$</td>
<td>31.02</td>
<td>32.14</td>
<td>48.19</td>
<td>61.31</td>
<td>53.95</td>
<td>67.29</td>
</tr>
</tbody>
</table>

Postman & Riley:

<table>
<thead>
<tr>
<th>Trial</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>40</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimated $k$</td>
<td>173,318</td>
<td>239,432</td>
<td>186,641</td>
<td>120,908</td>
</tr>
</tbody>
</table>

Rubin et al.:

<table>
<thead>
<tr>
<th>Trial</th>
<th>4</th>
<th>7</th>
<th>12</th>
<th>21</th>
<th>35</th>
<th>59</th>
<th>99</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimated $k$</td>
<td>37,187</td>
<td>49,732</td>
<td>45,602</td>
<td>55,711</td>
<td>53,603</td>
<td>53,551</td>
<td>41,075</td>
</tr>
</tbody>
</table>

Recognition experiments

Fajnsztejn-Pollack:

<table>
<thead>
<tr>
<th>Week</th>
<th>2</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>49</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimated $k$</td>
<td>0.96</td>
<td>0.69</td>
<td>0.46</td>
<td>0.52</td>
<td>0.32</td>
</tr>
</tbody>
</table>
Dilution as a model of forgetting

1 In referring to Jost’s law throughout this article we mean specifically his second law, concerned with forgetting.
2 In the analysis of recognition memory later in this paper, we apply principles of Signal Detection Theory to suggest that a trace population is further processed to manipulate the composition of trace populations identified by the available cues. Although this is not a necessary assumption for modelling cued recall, we note that these later models question the simplicity of assumptions made here for the purpose of exposition and to which we return.
3 We optimise the sums of squares of the natural logarithms of the probability of recall as a function of delay to reduce the dependence of variance upon mean. This is a particular issue in this case because it would weight parameter estimation towards data derived from the shortest delays. Quoted percentage variances as measures of model fit are calculated as proportion of the sum of squares of the model fit to the data divided by the sum of squared deviations of the data points about their mean. Throughout, the term log-log transformation refers to natural logarithms.
4 The data are estimated from Slamecka and MacElree (1983, Figure 4), being otherwise unavailable.
5 This is not to say that other processes not discussed in this paper may not have an effect upon the observed retention functions beyond any predictions of the PD model. Here, although there are insufficient data points at longer delays to test any such claims robustly, there is a suggestion in the untested data (U1 and U3) that the retention functions asymptote at a level greater than 0, as indicated by these points in Figure 6 lying above the theoretical predictions. Such an outcome could reflect a consolidation of a subset of the stimuli (e.g., see Wixted 2004), or it could reflect a residual probability of correct guessing. Estimating the latter by adding a guessing parameter increases the fit of the model marginally ($R^2 = .974$) with an estimate of a guessing probability of .103; which is plausible if participants have remembered a small number of target words and use them as responses when cues fail to elicit recall. None of the other experiments analysed in this paper are likely to be significantly affected by high guessing probabilities or other reasons for non-zero asymptotes. Indeed, Wixted (2004) illustrates well how misleading the assumption of an non-zero asymptote can be. Nevertheless, it is useful to observe that the PD model is not exclusive of the possibility. The effectiveness of the PD model across the experiments analysed in this paper suggests that their additional effect upon the retention function will be of a second order.
6 Collateral support distinguishing and excluding short term recall in this experiment is to be found in Rubin et al.’s reporting of reaction times (Figure 6, p. 1168), that shows shorter RTs for low lag values, but relatively stable values beyond lag 4.
7 Note in this case that the linkage between mean overall recall for a group and C is complex. The overall mean (a measure derived across all lags) is used opportunistically to differentiate the level of learning of the quintiles, and the hypothetical quantity $C$ is an estimate of the number of correct traces at lag 0 only.
8 The reason for assuming an additional decision process will be seen more readily when discussing the old-new paradigm in Fajnsztejn-Pollack’s and Rubin et al.’s recognition experiments below. This is necessary to maintain the internal coherence of the PD model and is included here for generality despite the fact that the parameters $P_0$ and $P_s$ measuring the proportion of supra-threshold traces are not estimable in this experiment; being subsumed in the parameter $k$.
9 This could also be construed as an old-new design. We adopt this approach to avoid introducing additional model constraints at this stage (for which this experiment is not an ideal vehicle) and because our principal purpose in analysing Strong’s data is to discuss meta-theoretic issues about how such data are interpreted.
10 The introduction of this additional process to model short term recall is inconsistent with our earlier explicit strategy of not complicating our models of the recall function with additional models for short term memory. We cautiously introduce it here because Figure 10 allows us to illustrate how misleading linear plots of data may be in the absence of supportive theory. Our sigmoidal function may appear less parsimonious than the log-linear plot (and is, given that it requires an additional free parameter) but the putative linear relationship between percent recognized and log time implied an atheoretical mathematical relationship that encouraged, but then defied, principled explanation (see also Köhler (1943) arguing for more theoretical analysis to avoid a premature reliance upon mathematical description alone).
11 The contingency of showing single pictures in this recognition paradigm requires the assumption of an additional decision process in modelling recognition memory. If, as in cued recall, we assumed that selection of responses was just the selection of one of the accessed traces, the absence of competing recognition cues would result in a false alarm probability of 1. Some additional mechanism is
Dilution as a model of forgetting

necessary in old-new paradigms to permit participants to judge the familiarity or strength of the foil. For generality, we have applied this mechanism throughout our analyses of recognition experiments. 12 The model expressed in Equation 3 for the likelihood of correct recognition in the 2AFC paradigm appears inconsistent with that for the old-new paradigm here, and some further comment is useful. Specifically, in the latter case, the entire population of traces activated by the recognition cue, including those of sub-criterion strength λ, are used to define the population from which a single trace is selected on a one-shot basis. In the 2AFC model however, sub-criterion traces (i.e., those of sufficient strength to be activated by the recognition cues(2) but of strength < λ) are excluded from selection; which is made only from traces exceeding the criterion strength. The simplest psychological interpretation of this is that memory is repeatedly sampled until a supra-threshold trace is revealed; which occurs with the likelihood of correctness given by Equation 3. This violation of the one-shot principle is justifiable on the grounds that the participant must select one or other of the recognition cues as a response (that is, saying ‘neither’ is not an option) and can be expected to seek evidence in memory to support their choice. This contrasts with old-new tests, where the failure to elicit a supra-threshold at first retrieval could be taken as evidence of a new item. We note that other more complex models of the 2AFC experiment are possible within the PDE model. For example, participants may compare the strengths of two traces retrieved on a one-shot basis from populations determined by both the target and foil. These vary in strength according to the distributions illustrated in Figure 8 and may be amenable to signal detection approaches in which the strength of traces is compared. However, implementation of such a model is not straightforward given the bimodality and numerical heterogeneity of the distributions. Given that the aim of this article is to demonstrate the general applicability of the principle of population dilution to forgetting across a range of experimental methodologies, we have preferred the simpler formulation here; leaving a detailed examination of the merits of different implementations of the PDE model as an issue to be considered subsequently.

13 This simulation is implemented, along with all other models presented in this paper, in an Excel spreadsheet and available at www/to.be.announced/

14 We use k in this case to differentiate the effects on recall of the time lag between the end of learning trials and the beginning of relearning, where the constant k is applied. This device also reinforces the idea that within the PD model ‘forgetting’ is taking place during the learning cycle (see also Sikström 2000), since the dynamics of retrieval in forgetting and learning stages of experiments are exactly the same.

15 Rubin et al. used 13 free parameters to model all 45 data points of the data illustrated in their Table 1. Note, however, that Rubin et al. were modelling the entire data, including short term lags. In this implementation, by discarding short term data, 35 data points are modelled with 4 free parameters.

16 All estimates discussed here have been derived from data aggregated across participants. We recognise that the possibility that the mathematical form of the aggregate model differs from that at the level of individual participants (e.g., see Sikström, 2002; Wixted, 2004) cannot be ruled out, and requires further investigation.

17 In the cases of Postman and Riley and Rubin, Hinton and Wenzel’s two experiments, where the units of delay are trials, we have conservatively assumed for Table 2 that each trial took one minute. The effect of this is to underestimate D, since this time was probably considerably shorter. However, since our conclusions rest upon observed high values in these cases, this is not problematic.

18 Although it is very tempting to interpret k as simple parameter quantifying the amount of interference in the experiment, this interpretation is misleading. As k is a ratio, it is very sensitive to the base rate of the denominator. For example, as the number of correct traces approaches zero the denominator of the ratio approaches zero and k tends to infinity. The difficulty in interpreting k is analogous to that of an odds ratio. An odds ratio is a convenient general measure of effect size between studies but needs to be coupled with base rate information in order to make useful inferences about a specific situation.

19 Note also that the sensitivity of these estimates in the PD model varies as a function of n. Subsidiary analysis, not reported here, demonstrates that perturbing estimates of Cs by fixed percentages of the optimised value decreases the goodness-of-fit of the model as n falls. With higher values of n, the overall trace population in the PD model is larger, and therefore the same absolute variation in Cs, the sole numerator in Equation [1], has less impact on the predicted retention function.

20 The distinction between short and long term memory can be made on different grounds. Strong arguments have been made that experimental support for the distinction is actually lacking (e.g., Laming, 1999). Even so, short and long term memory experiments can also be distinguished by the different constraints they place upon the strategies and behaviours available to the participants, with
Dilution as a model of forgetting

various processes of rehearsal and temporal discrimination available in the short term not available in the long.
[21] Which interestingly, given our suggestion of null traces, also includes the concept of ‘junk’ memories.