We'll remember it for you wholesale

Simon D. Lilburn

27 May 2013

Funes, the Memorious

> Without effort, [Ireneo Funes] had learned English, French, Portuguese, Latin. I suspect, nevertheless, that he was not very capable of thought. To think is to forget a difference, to generalize, to abstract. In the overly replete world of Funes there were nothing but details, almost contiguous details.
>
> (—*Funes, the Memorious*, Borges 1962)

The Argentine writer Jorge Luis Borges wrote short stories, essays, and poems effusive with insight into the construction and perception of the world. Borges was a writer who examined philosophy, psychology, and logic in the way that a science fiction writer might examine physics, astronomy, or computing.

Borges's Ireneo Funes lives in a world incommensurate with our own experience. His past is as vital and urgent as our present. Where we must endure a time for coloured by the generalisation and change and obliteration of memory and experience, he must endure unabstracted, unsynthesised singularites.

Funes is a tragic character: dispossessed of thought and held hostage by a history which is at once inert and oppressive.

The failure of memory

Forgetting

The principal characteristic of human memory is a lack of fidelity. The most salient of these corruptions is forgetting. The uniformity in how accuracy in tasks designed to probe human memory falls away over time—and the quantification of this uniformity—has served as one of the elemental results of modern experimental psychology.

Ebbinghaus[1] first described the functional form of forgetting using the learning, recitation, and relearning of lists of three letter nonsense syllables (called CVC trigrams to reflect their composition: consonant–vowel–consonant) over extended time periods. The savings gained on relearning a previously learned list at different intervals (as a quotient of the original time to fully learn the list) was used to infer the structure of the "oblivescence" (forgetting) of memory traces.

The constancy and detail of these findings with differing paradigms and stimuli has provided an enduring line of theoretical discussion[2]. Recent work by Donkin and Nosofsky[3] demonstrates the existence of power-law forgetting curves (of the form $m = \alpha \times j^{-\beta}$)

[1] Ebbinghaus, H. (1885). *Über das Gedächtnis. Untersuchungen zur experimentellen Psychologie.* Duncker & Humblot, Leipzig

[2] Wixted, J. T. and Ebbesen, E. B. (1991). On the form of forgetting. *Psychological Science*, 2(6)

[3] Donkin, C. and Nosofsky, R. M. (2012). A power-law model of psychological memory strength in short-term and long-term recognition. *Psychological Science*, 23(6):625–634

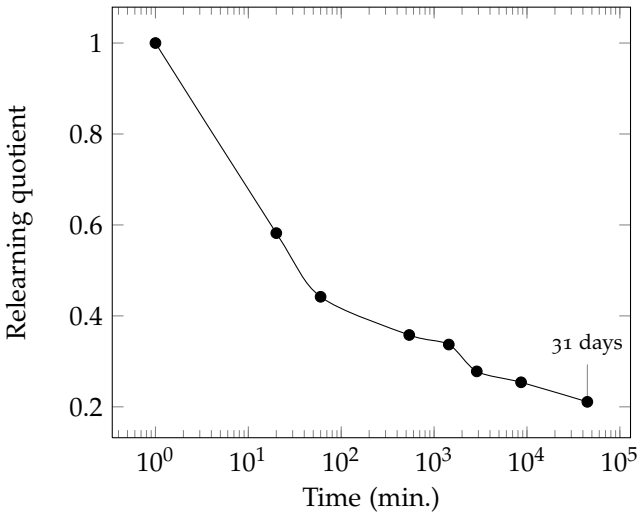

Figure 1: The "forgetting" curves of Ebbinghaus (1885). Forgetting is operationalised as the proportion of the original memory list that must be relearned to return to the peak initial performance. Note the approximate linearity of the function over the logarithmic scale of time.

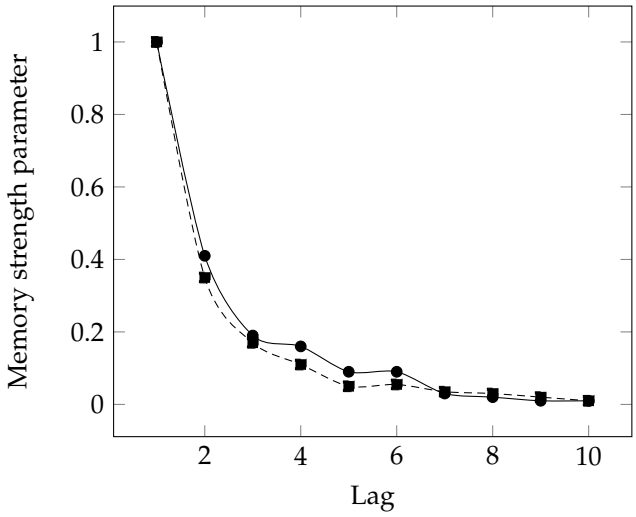

Figure 2: Donkin and Nosofsky (2012) found evidence for power-law "forgetting" curves in the sensitivity parameters of an exemplar–recognition model (specifically, the exemplar-based random walk model of Nosofsky and Palmeri) in both short- and long-term recognition memory data. The "lag" measure in the figure denotes the serial position of the target item in a sequentially presented memory array with respect to the probe. Different lines on the chart represent different parameters for two participants.

in short- and long-term recognition memory paradigms (experimental paradigms where memory is assayed through the use of "probe" items which may or may not have been on a previously studied list).

False memory

ELISIONS ARE ONLY ONE PART OF MEMORY FAILURE. Bartlett[4] demonstrated that individuals reciting a previously learned story will confect—as well as omit—details of the story in subsequent retellings. Based on the patterns of change between the original story and the retold versions, Bartlett argued that the mode of retrieving information from memory was necessarily active and reconstructive: elements and details were extracted from expectations codified in memory schemata.

[4] Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge University Press, Cambridge, England

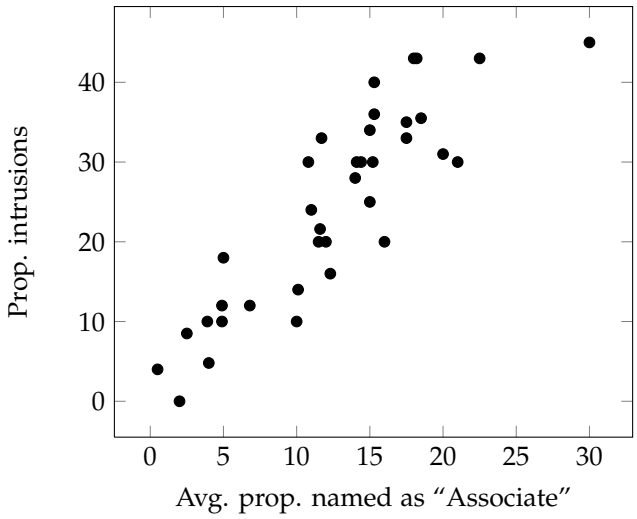

Figure 3: Deese (1959) demonstrated an association between list "intrusions" (items that are falsely recalled as being on a given memory list) and the proportion of judgements indicating that a word is similiar (an associate) of the words on the memory list.

These results have been elusive to replicate and further characterise due to the indeterminacy of novel responses in recall tasks (*viz.* it is difficult to say why any detail might be absent, might appear, or might change between individuals or within multiple recitations in a recall paradigm). Deese[5] provides one instance where experimental data from a free recall paradigm can demonstrate consistent patterns of "false recollection" content: words are more likely to be falsely recalled if they are judged to be more similar to items that appear on the memory list.

[5] Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, 58(1):17–22

Recognition memory paradigms provide more tractable theoretical questions due to the constraints on the response type (often participants will be asked simply to indicate whether item is "old" or "new" or rate their familiarity with an item on a given scale). In "old"–"new" recognition memory tasks, each response can be classified in one of four ways: a "hit" (correct identification that an item was previously studied); a "miss" (a previously studied item dismissed as novel incorrectly); a "false alarm" (a novel item endorsed as previously studied incorrectly); and a "correct rejection" (a novel

item correctly identified as such). The precise categorisation of responses into these categories allows strong constraints to be placed on the underlying memory process.

The presence of "false alarms" in recognition tasks is analogous to "false recollections" in recall tasks, a novel item incorrectly identified as having been seen before.

Signal detection theory

Classical threshold theory

Researchers contemporary with Ebbinghaus—namely Weber, Fechner, and Wundt—were providing the first rigorous description of the transduction from physical phenomenon to elicited perceptual experience. The data from these investigations formed the basis for a continuing research programme in experimental psychology: psychophysics.

Magnitude was the first subject to be examined: the first inquiries in psychophysics investigated the translation between physically realised quantities of (brightness, sound pressure level, pressure) to perceptual intensities.

At the lowest perceivable intensity levels, at the threshold of perception, it might be intuitive to conjecture an all-or-none sensory *limen*: this would predict that people are unable to report having perceived a stimulus below a certain physical intensity until some critical value after which they can consistently report the presence of the stimulus. Across observers and sensory modalities, this was found not to be case. Rather, a sigmoidal ("S"-shaped) curve was found around the threshold.

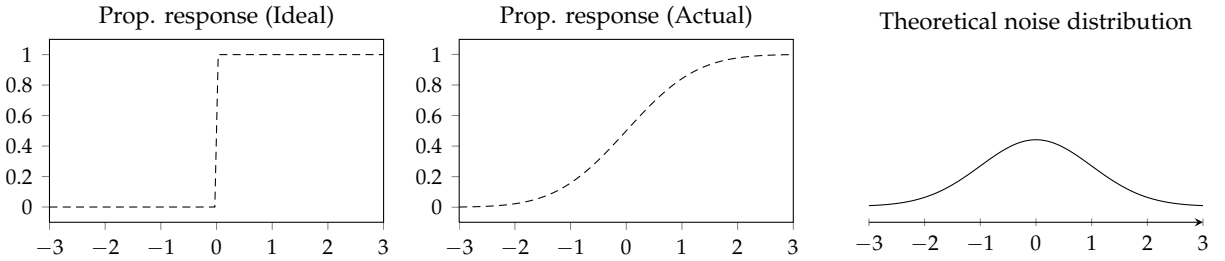

Figure 4: Where one might expect a hard threshold at the lowest levels of intensity detected by human sensory systems, a sigmoidal curve is usually found. This smooth function is taken to indicate a distribution of noise that perturbs the representation during mental processing.

This smooth ("psychometric") function of responses made from perceptual representations over physical levels of intensity was taken to indicate some background (neural) noise in the processing of sensory representations, which might perturb and degrade the representations. This means that, on some proportion of presentations where the stimulus is above a theoretical threshold, noise would cause the stimulus to be missed and, likewise, on some proportion of presentations where the stimulus is below a theoretical threshold, noise would cause the stimulus to be registered.

Signal detection theory

A PROBLEM FOR CLASSICAL THRESHOLD THEORY is that instructions to individuals performing sensory detection experiments (that is, experiments asking participants to respond upon perceiving a at-threshold event occurring) can the location of the mean sensory threshold along the continuum of physical intensities. In particular, it was found that when instructed to be "strict" participants would be more conservative about their judgements regarding stimulus detection and vice versa when instructed to be "lax" about detection responses[6]. This is not predicted by classical threshold theory, which assumes that a there is a fixed underlying threshold which must be cleared by a signal (perturbed by noise).

The theory of signal detection[7], or signal detection theory, provided a parsimonious account of why such a phenomenon would occur. As in classical threshold theory, noise pervades the sensory systems causing fluctuations in the sensory signal. Often, however, it is the case that the strength of a signal (say, in the case of low intensity stimuli or vaguely recollected memories) will not fully clear values that uninformative noise can take. Thus, in some cases it will be ambiguous whether a level of activation has been reached by some low stimulus intensity value or some high noise intensity value.

Rather than a fixed threshold, which divides the continuum of intensity into a detectable and an undetectable region, it may be the case that—in order to respond in an adaptive manner—the boundary (or criterion) at which the response changes (from "detected" to "undetected" or from "old" to "new" in a recognition task) is not a hard property of the system, but tunable to ensure that behaviour is well suited to the task or environment at hand.

Put another way, given some ambiguous region, a trade off must be made: if all of the signals must be correctly identified by the system (i.e., no "misses") then the false alarm rate must increase as the intensity value at which most of the signal distribution is given a "signal" response includes a higher proportion of the noise distribution. Conversely, if false alarms are to be avoided, then the number of misses must increase as to keep the majority of the noise distribution from being false identified as a signal. A tunable criterion means that this trade-off can be made by the system (a perceptual or memory system) to the statistical properties at hand.

This notion of overlapping signal and noise distributions and a tunable criterion (able to tuned for task requirements or expectations) is at the heart of signal detection theory.

These results also have implications for memory systems. Since Egan[8], recognition memory has been described in signal detection theory terms. Dunn[9] and others have argued that there is strong evidence that the "remember"–"know" paradigm previously argued to account for two incommensurate systems can be fully

[6] Swets, J. A., Tanner, W. P., and Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review*, 61(5):301–340

[7] Peterson, W. W., Birdsall, T. G., and Fox, W. C. (1954). The theory of signal detectability. In *Proceedings of IRE Professional Group on Information Theory*

[8] Egan, J. P. (1958). Recognition memory and the operating characteristic (Tech. Note AFCRC-TN-58-51). Technical report, Hearing and Communication Laboratory, Indiana University

[9] Dunn, J. C. (2004). Remember–know: A matter of confidence. *Psychological Review*, 211(2):524–542

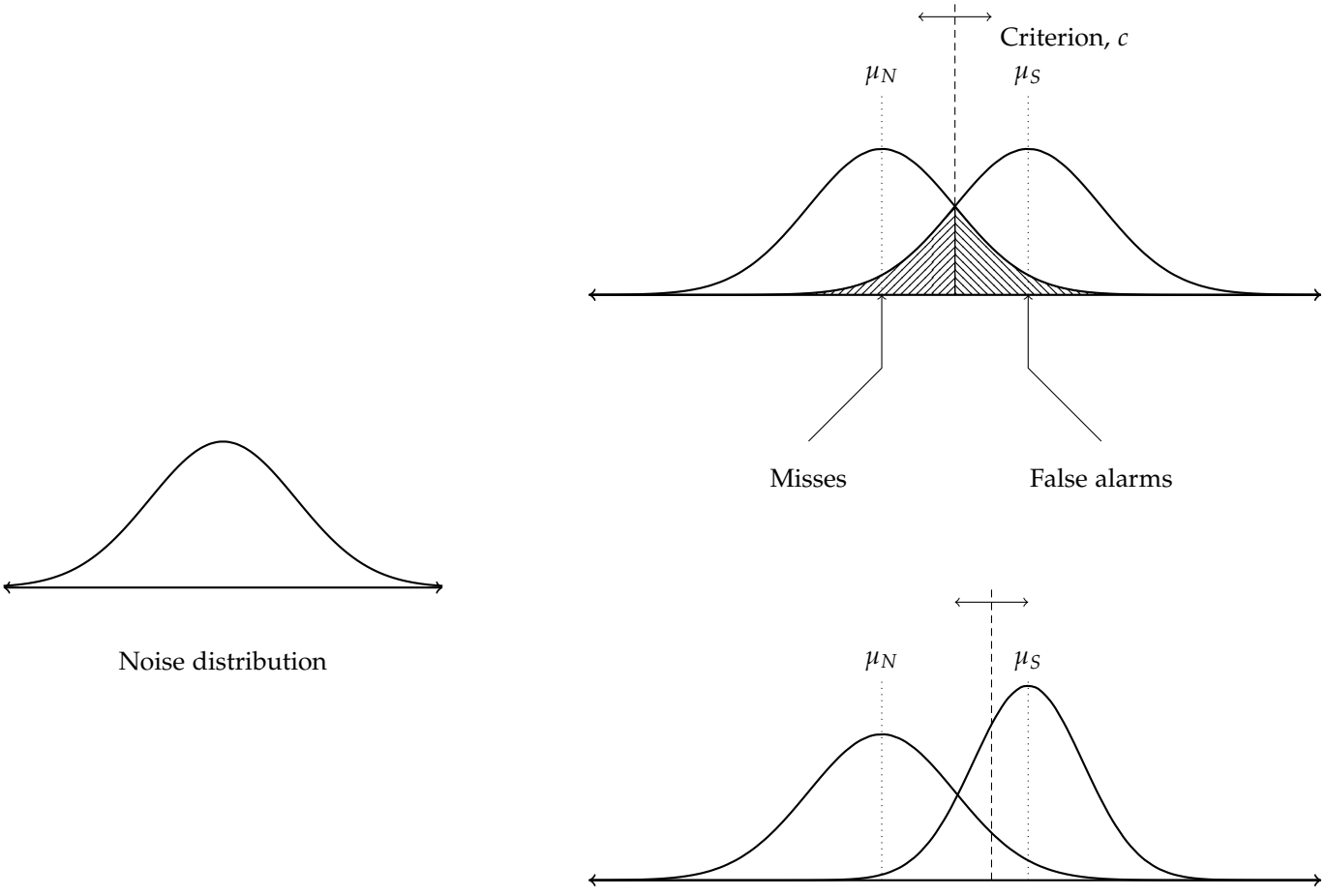

Figure 5: A schematic of the major elements of signal detection theory. A theoretical noise distribution (left) perturbs an additive signal (top right panel) or a signal with multiplicative interaction (bottom) leading to equal or unequal variance, respectively. The criterion can be set across the stimulus dimension, leading to different proportions of misses and false alarms. Note that where the height of the two distributions where the criterion intersects can be used to compute the likelihood ratio of the response.

accounted for by a criterion shift.

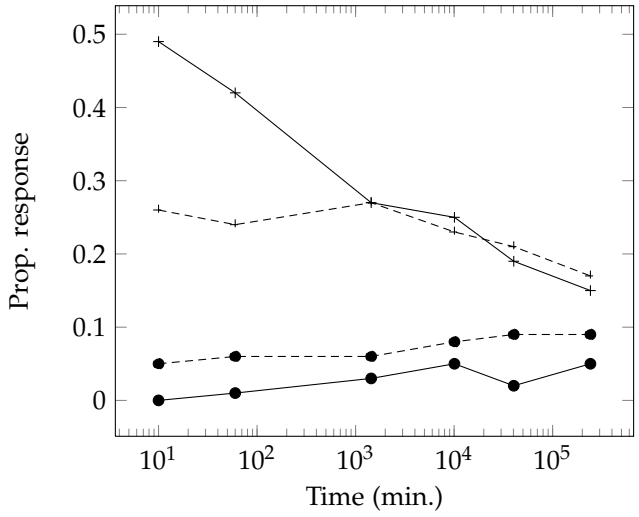

Figure 6: Data from Gardiner and Java (1991) examining hits and false alarm rates in the "remember"–"know" paradigm. Solid lines are from the "remember" condition; dashed lines are from the "know" condition. Plus marks represent hit rates; dot marks represent false alarm rates.

Four empirical features of recognition memory

SCIENCE IS SERVED WELL BY THE FORMALISATION OF IDEAS INTO RIGID QUANTITATIVE RELATIONSHIPS. Psychology gains especial benefit from this process of formalisation as many of the key ideas and relationships in psychology are borne of common experience and intuition. Although human experience might be the basis for psychological investigation, it is not necessary for phenomenological descriptions of mental processes to be clear, precise, and accurate enough to provide meaningful insights.

Any quantitative model of recognition memory—and potentially of memory at large—must first align with empirical evidence consistently demonstrated. The following four effects have proven to both be consistent across experiments and difficult to explain in terms of quantitative memory modelling.

The list length effect

IT HAS BEEN CONSISTENTLY FOUND in both recall and recognition tasks that increasing the number of items in memory list decreases the average sensitivity of any of the items. This is known as the *list length* effect. The first demonstration of this effect was given by Strong[10] who found that increasing the number of advertisements in a to-be-remembered list decreased mean recognition of any item later on.

Subsequent experiments[11] examining the list length effect—and analysis in terms of signal detection theory—indicates that list length can be described in terms of a decrease in sensitivity (that is, a decrease in d' which decreases both the "hit" rate and

[10] Strong, E. K. (1912). The effect of length of series upon recognition memory. *Psychological Review*, 19(6):447–462

[11] Gillund, G. and Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, 91(1):1–67

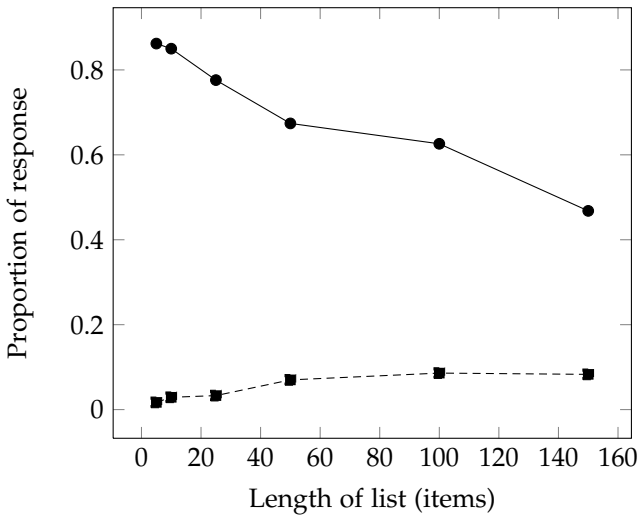

Figure 7: Strong (1912) found that increasing the number of advertisements in a memory list decreased the proportion of items correctly recognised during the test phase of the experiment. Note also that the number of incorrect recognitions also increases as the length of the memory list increases. The solid line denotes correct responses; the dashed line denotes errors.

increases the "false alarm" rate). The proposed cause of this effect is an increase in the dispersion of distractor/noise distribution, meaning that the strength of the (familiarity) signal decreases as a ratio of the noise and a greater portion of the signal distribution overlaps with possible noise values.

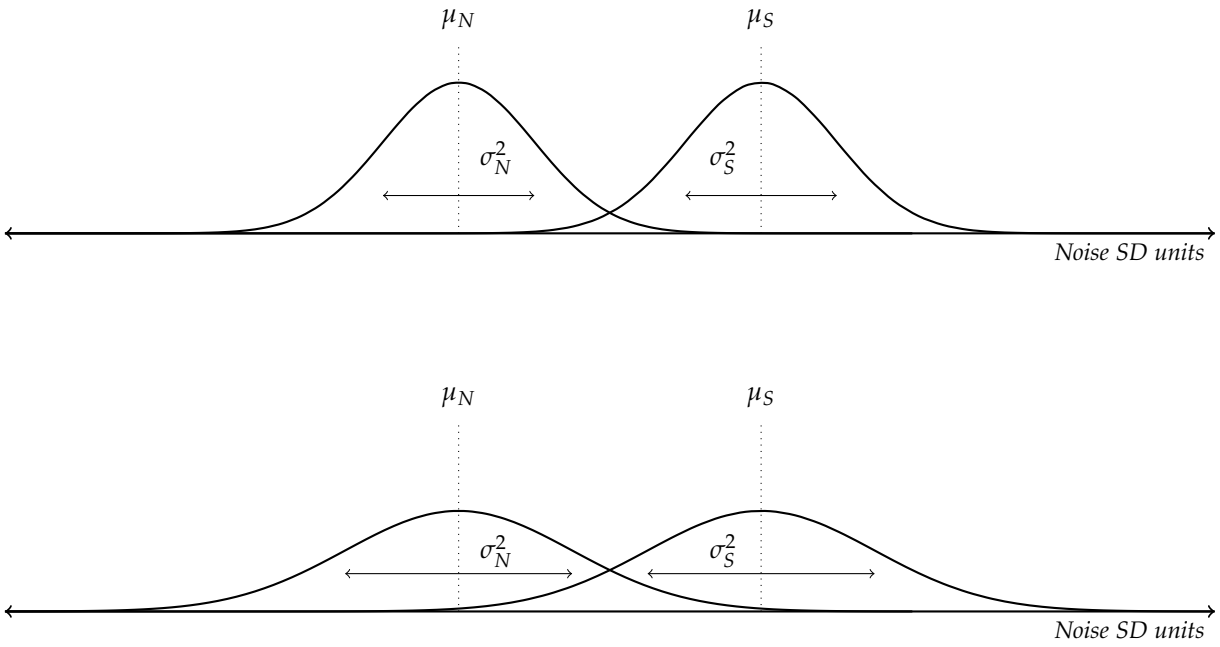

Figure 8: The proposed theoretical account for the list length effect in signal detection theory terms. The SDT measure of sensitivity, d', is the distance between the means of the noise and the signal distributions scaled in terms of the noise dispersion (standard deviation units). As the list length increases, the variability of the non-target (distractor) items also increases indicating that any fixed signal strength will decrease in proportion to the dispersion of the noise distribution.

The strength effect, the list strength effect

It has been demonstrated consistently in both recall and recognition tasks that it is possible to experimentally manipulate the probability of later correctly retrieving a stimulus in the memory array by manipulating the encoding/presentation conditions, usually increasing the time of initial presentation and encoding or repeating the stimulus multiple times. This is known as the "strength" effect, indicating the underlying memory trace is "strengthened" and resistant to decay over time or interference from other items.

It is reasonable to assume—analogous to the explanation that additional distractor items cause greater variability in non-target (non-signal) information—that strengthened items within a list would have a similar effect. That is, when strong items are present in a memory list with weaker items, the sensitivity of the weaker items would be compromised. This theoretical effect—predicted by many models of memory performance—is known as the *list strength* effect.

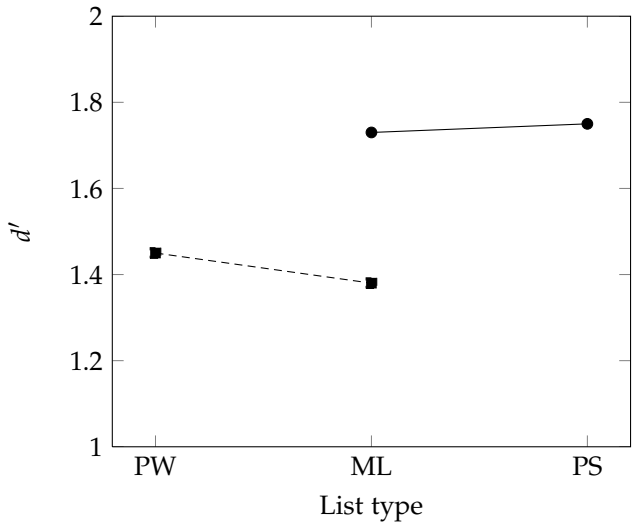

Figure 9: Data from Ratcliff et al. (1990). There is no indication of a major and systematic effect of mixing item strengths causing a list strength effect in recognition paradigms. This is contrary to many models assumptions of the effect of strength on distractor variability.

List types: PW = "Pure weak"; ML = "Mixed list"; PS = "Pure strong".

This effect does present empirically in recognition tasks, however. Ratcliff, Clark, and Shiffrin[12] demonstrated that "mixed" lists composed of items presented to ensure stronger or weaker encoding did not greatly differ in their sensitivity when compared to the same items in "pure" weak-only or strong-only lists. This is a difficult result to explain when memory models assume that item strength would have an analogous effect to list length.

[12] Ratcliff, R., Clark, S. E., and Shiffrin, R. M. (1990). List-strength effect: I. data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(2):163–178

z-ROC profile

Receiver operating characteristic (ROC) curves quickly describe the profile of sensitivity in a detection system by plotting the proportion of correct identifications of a signal (in our case, a

previously studied stimulus) against the proportion of incorrect identifications of noise (distractors) as being as signals (items from the memory array).

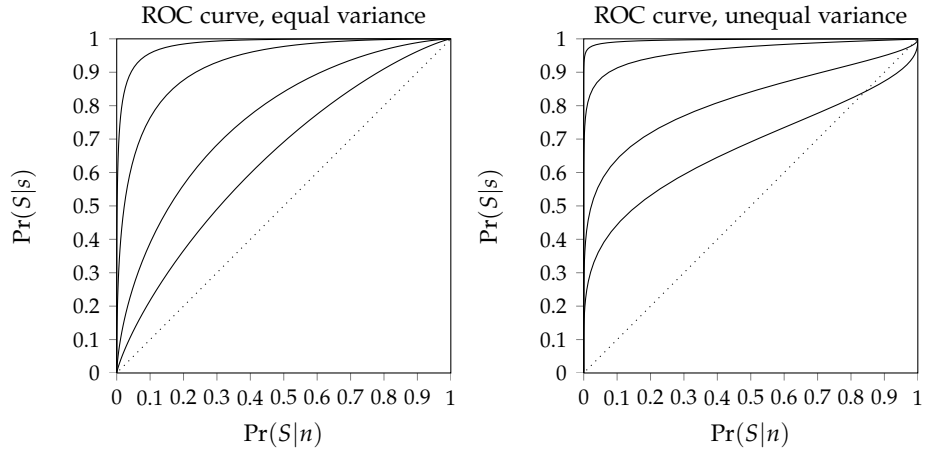

Figure 10: ROC curves representing detection sensitivity where the noise and signal variance are either equal or unequal (greater signal variance). The value $\Pr(S|n)$ represents the false alarm rate (the chance of responding signal S given the presentation of noise n); likewise, the value $\Pr(S|s)$ represents the hit rate (the chance of responding signal S given the presentation of a signal s).

Each line represents a different value of d', with the lines closer to diagonal representing lower sensitivity values.

These curves not only give insight in to exactly how sensitive a system can be with any given criterion (that is, how many hits and false alarms might be incurred with a shift in the response criterion), but also the shape of the noise and signal distributions. Conversion of ROC curves into "normal" or z space (that is, a coordinate system based on the standard normal distribution) functions like a Q-Q plot, where linearity corresponds to a distributional form shared between the false alarm and hit rates (and, therefore, between the signal and noise distributions).

In an extended examination of the properties of ROC curves in recognition memory paradigms, Ratcliff, Sheu, and Gronlund[13] found that the ROC curves in recognition memory tasks show a remarkable stability. Characteristically, the ROC curves indicated that the signal distribution had the same shape as the noise distribution, but a consistently larger variance.

[13] Ratcliff, R., Sheu, C.-F., and Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review*, 99(3):518–535

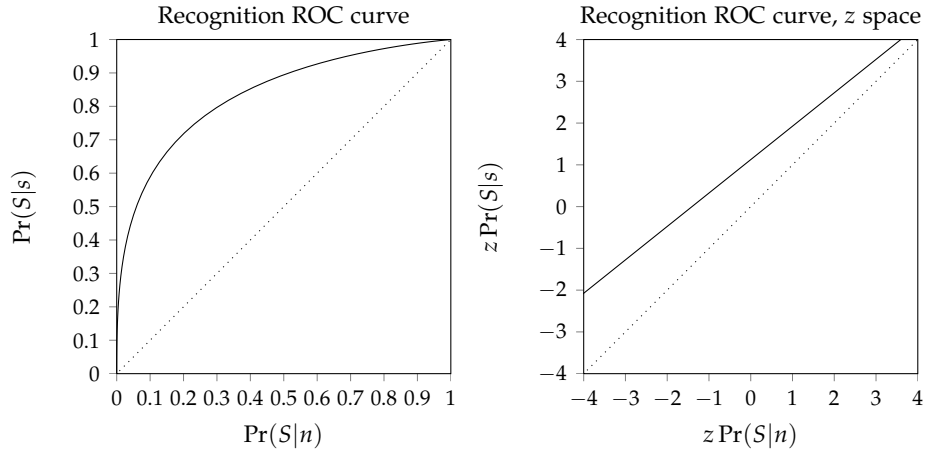

Figure 11: ROC curves found by Ratcliff et al. (1992) in recognition memory tasks. Note that, in z-space, the ROC curve is a linear function with a slope of 0.8. This indicates that the signal distribution has the same shape as the noise distribution, but a larger variance.

The word frequency effect and the mirror effect

LINGUISTIC STIMULI ARE OFTEN USED IN RECOGNITION MEMORY EXPERIMENTS because the statistical properties of words can be accurately described and quantified against large corpora of texts, which is difficult with most other kinds of stimuli that would be employed in a recognition memory task. The frequency of words occurring in the larger environment, for instance, can be approximated by looking at their occurrence in a large and representative set of words.

Word frequency produces consistent empirical effects in recognition memory tasks: words with a higher frequency are less likely to be correctly recognised and more likely to cause higher false alarm rates than words with lower frequencies. This word frequency effect indicates that high frequency words have a lower d' than low frequency words: the strength of a recognition signal is smaller as a proportion of the noise variance between high frequency words when compared to low frequency words.

A second, more subtle, effect of word frequency (and other effects of linguistic attributes on sensitivity) was demonstrated by Glanzer and Adams[14]: the mirror effect. The mirror effect gives rise to a specific order of hit rates and false alarm rates that imply an ordering not implied by signal detection theory. This order of distributions is shown in Figure 12.

[14] Glanzer, M. and Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition*, 13(1):8–20

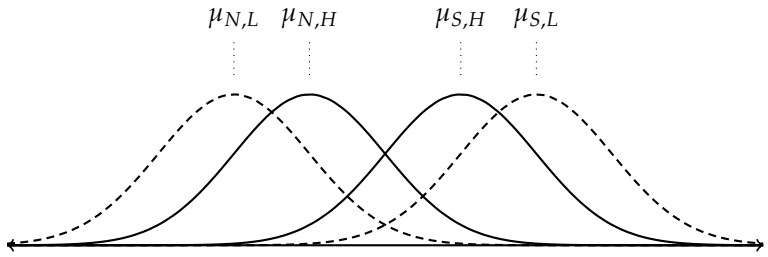

Figure 12: The ordering of the underlying theoretical distributions required to give a mirror effect (Glanzer and Adams, 1985). The dashed distributions represent the signal and noise distributions for low frequency words; the solid distributions represent the signal and noise distributions for high frequency words. The means are labelled above the distribution, with S denoting signal distributions, N denoting the noise distributions, L denoting the distributions for low frequency words, and H denoting the distributions for high frequency words (e.g., $\mu_{S,H}$ is the signal distribution of high frequency words).

Theoretical building blocks: features and likelihoods

Features

THE FIRST CHALLENGE FOR ANY MEMORY MODEL is the quantify the nature of the underlying memory representation and, then, how the prediction about those representations will interact with the empirical results to explain. Like all psychological phenomena, the constitution of a memory representation is not immediately apparent from experience.

One incontrovertible assumption is that—in order for human beings to be adaptive—memory systems must some how reflect or be sensitive to the structure of the environment[15]. That is, in order for memory systems to have some functional use in an environment,

[15] Anderson, J. R. and Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, 96(4)

they must be able to perceive and represent some aspects about that environment—this is usually assumed to be the probability distributions for events and features occurring. The argument that cognitive science should examine the ways in which organisms and cognitive systems exist and operate adaptively within their environments is known as "rational analysis".

Although this assumption does not impose a particular structure on the representation (indeed, it can seem like something of a truism), it does suggest that a fruitful approach to examining memory is to examine how statistical properties of the environment might be effectively and efficiently represented in a memory system.

To return to Borges's Funes, we might say that the critical error of the Funesian memory is—as the narrator of the story notes—abstraction is not possible from memories that are stored as monadic wholes. Further, many characteristics in any one object are shared with other instances of its category. To store each of these characteristics in full or partial fidelity each time for each item when so many features would be shared between items would be grossly inefficient.

An alternative theoretical perspective is advanced by Bower[16]: rather than memory representations being stored as complete indivisible wholes, memory traces might be represented by strings of subunits of representation called "features". In this account, a feature corresponds to some way that stimuli vary in the environment. These means of variation might be declarative (e.g., a person might have hair or be bald) or unable to be easily rendered into words. The identity of each memory trace would be set by values of the features.

These findings are conducive with findings computational models designed to resemble the action of populations of neurons[17]. On a psychological level, an explanation of features immediately allow us to examine why the false alarms occur: two items might have representations differing by only a small number of features. If we assume that noise can cause features to be erronenously encoded or retrieved, then it seems likely that one representation might be confused with another. Bower also demonstrated that features could be used to explain memory phenomena like the forgetting curve in recall and recognition memory and repetition effects.

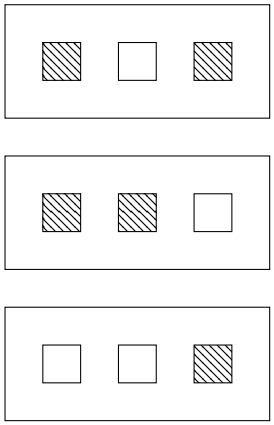

Figure 13: A visual representation of three items stored in memory, each with three binary features (where an unfilled square represents an "off" feature and a marked square represents an "on" feature). The top and the middle memory items only differ by a single feature value, as do the top and the bottom items. Noise in a representational system might cause confusions between the top item and the other two. The bottom two items are different by two features and might, therefore, be relatively unlikely to be confused with one another.

[16] Bower, G. H. (1967). A multicomponent theory of the memory trace. In *The psychology of learning and motivation: Advances in research and theory*, pages 229–325. Elsevier

[17] Rumelhart, D. E. and McClelland, J. L. (1986). *Parallel Distributed Processing, Vol. 1: Foundations*. MIT Press, Cambridge, MA

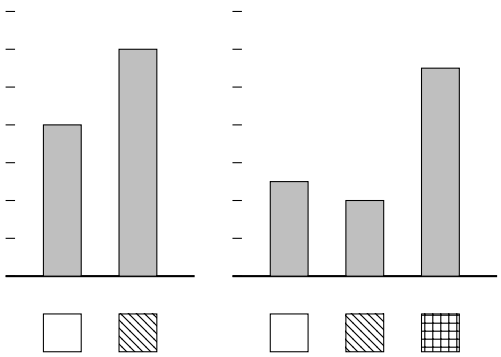

Figure 14: Features might have different types and distributions in the real world. For instance, the first panel displays a binary feature (that is, a feature that can only take two values). In the environment, the first feature value is slightly less probable than the second. The second panel displays a ternary feature; the third value of this feature is much more likely to be seen in environment than the first two values.

Probabilities, odds, and likelihoods

THE UTILITY IN DESCRIBING EVENTS AND FEATURES IN TERMS OF PROBABILITIES is that it is possible to precisely define what the relationship between occurrences of events and expectations should be. In terms of utility to the organism, it is clear why a precise account of the probability of events and features in the world would be a good thing: an organism performing optimally in an environment is unable to do better on average.

In order to examine the relationships between events or facts, we need to define the relationships between their probabilities. Conditional probabilities give us a way of talking about the relationship between of two probabilistic events: $\Pr(A|B)$, the probability of fact A being true given that fact B is true. The definition of conditional probability is

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

That is, the probability of A being true given that B is true is the ratio of both A and B being true over the probability of B occurring. This definition immediately allows us, by very minimal algebraic manipulation, to obtain a relationship between the two conditional probabilities $\Pr(A|B)$ and $\Pr(B|A)$ using the insight that the conjunctive term ($\Pr(A \cap B)$) is shared between them. Substituting one conditional probability in the equation for another gives us Bayes' theorem:

$$\Pr(A|B) = \frac{\Pr(B|A) \cdot \Pr(A)}{\Pr(B)}$$

The utility of Bayes' theorem becomes evident when we allow probabilities to represent not only observable events occurring in the world, but also our beliefs in facts being true about the world. That is, if we can construct hypotheses about how events in the world might be produced, and give probabilities to those hypotheses being true, then we can use the model, the data, and our belief in the model to indicate strength of support the data have for a model. Formally,

$$\Pr(\mathcal{M}|D) = \frac{\Pr(D|\mathcal{M}) \cdot \Pr(\mathcal{M})}{\Pr(D)}$$

where D represents the data, $\mathcal{M}$ represents the model which can produce the data. Each of the terms has a specific name and role: the prior distribution, $\Pr(\mathcal{M})$, represents our degree of belief in the model (the probability of model being true); the posterior distribution, $\Pr(\mathcal{M}|D)$, represents the updated degree of belief in the model given the data; the quotient $\Pr(D|\mathcal{M}) / \Pr(D)$ represents the strength of the evidence the data provides for the model. This process of updating our beliefs in models based on prior beliefs and data is called Bayesian inference.

The conditional probability of the data on the model, expressed by the term $\Pr(D|\mathcal{M})$, is known as the likelihood. This probability

represents the probability that the model predicts for the data being true (technically, the probability of the data occurring given that the model is known to be true). The likelihood represents the key step in the Bayesian inference: it is the direct measure of how well the hypothetical model accounts for the actual data.

The likelihood also serves another important role. Given two models both able to predict a single set of data, we can compute the ratio of the likelihoods to quantify the relative difference between the two models in accounting for the data. By taking the expression for Bayes' rule with two models, $\mathcal{M}_1$ and $\mathcal{M}_2$, we obtain

$$\frac{\Pr(\mathcal{M}_1|D)}{\Pr(\mathcal{M}_2|D)} = \frac{\Pr(D|\mathcal{M}_1)}{\Pr(D|\mathcal{M}_2)} \cdot \frac{P_O(\mathcal{M}_1)}{P_O(\mathcal{M}_2)}$$

where the term $P_O(\mathcal{M}_1)/P_O(\mathcal{M}_2)$ represents the "prior odds" of how probable each model is to be true, the term $\Pr(\mathcal{M}_1|D)/\Pr(\mathcal{M}_2|D)$ represents the likelihood ratio of the two models given the data, and the term $\Pr(\mathcal{M}_1|D)/\Pr(\mathcal{M}_2|D)$ represents the odds after the data. In the case where the prior odds favour no model, the posterior odds give the relative weight that should be apportioned to each of the models given the dataset.

An example: the mirror effect and signal detection theory

SIGNAL DETECTION THEORY encapsulates the notion of decision making using likelihood ratios.

If a stimulus is presented at a level which is ambiguous (that is, it could have been drawn from the noise distribution or the signal distribution), the probability of both noise or signal is non-zero at the criterion. The criterion is often expressed in terms of β, a likelihood ratio of the noise and signal distributions at a point of the response criterion,

$$\beta = \frac{y_S}{y_N}$$

where y is the height of the distribution at the criterion (the ordinate), S is the signal distribution, and N is the noise distribution.

If both distributions are at equal height at the criterion, then β will be equal to one. This means that, given that if there is an overlap between the signal and the noise distributions there must be a trade-off between hits and false alarm rates, the trade-off is unbiased in terms of favouring correct responses (hits or correct rejections) for one distribution over another. Put another way, assuming normal distributions, this means that the mean of the signal distribution is the same distance away from criterion as the mean of the noise distribution.

The commonly accepted explanation for the mirror effect is that it is often optimal to make recognition judgements with an unbiased criterion, a likelihood ratio of one. If the sensitivity in one condition is lower than the sensitivity in another, but the likelihood ratios are the same, then a mirror effect is necessarily obtained.

REM

The REM (Retrieving Effectively from Memory) model of Shiffrin and Steyvers[18] amalgamates many of these ideas into a single, simple model.

[18] Shiffrin, R. M. and Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, 4(2):145–166

Précis

The REM model is a straightforward formalisation of three widely held assumptions: first, that representations within memory (termed "images") are composed of conjunctions of independent features; second, that those features have distributions in the environment, and that the memory system has internalised the likelihood of any feature value occurring; and third, that the ratio of the likelihood that a probe has been previously seen and the likelihood that a probe has been randomly sampled from the environment can be computed.

Representations and encoding

In REM, memory representations are composed of independent features. A single representation in memory (an "image" or "trace") is composed of a set of these features set to individual values. Each feature can take a value drawn from a geometric distribution with some rate parameter, g. That is,

$$\Pr(V = j) = (1 - g)^{j-1} \cdot g, \quad j = 1, \ldots, \infty$$

where V is the feature value and g is the rate parameter. The geometric distribution has the implication that, as the rate parameter gets higher, fewer values are seen in the environment. Likewise, with a low rate parameter, more probability mass is in the long tail, meaning a larger range of values that will be seen on average.

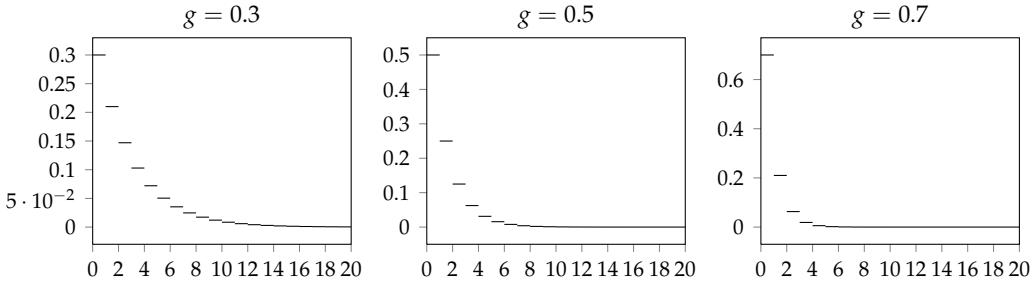

Figure 15: Probability mass functions (pmfs) for a geometric distribution over the first twenty natural numbers, with g set to 0.7, 0.5, and 0.3.

As features are constrained to independently vary in the environment, the probability of any image (assuming images are equally probable to be sampled from environment) is simply the product of the density of each features,

$$\Pr(\mathbf{X}) = \prod_{i=1}^{n} \Pr(x_i)$$

where $\mathbf{X}$ is the full feature set (image) and n is the number of features.

When items from the study list are being encoded, REM assumes that the complete image is presented to the memory system for encoding. This encoding happens in discrete time steps. Within each given time step, unencoded features have some independent probability, u^*, of being encoded. The number of time steps is manipulated experimentally as the length of the presentation time. If, after all of the time steps, a value is left unencoded, it is given a blank "0" value and will be uninformative in future comparisons (the feature value gives no clue to what the original value was).

It is also assumed that encoding is noisy and that there is some chance that a feature will be misrepresented in the stored memory image. If a feature is being stored on this step, there is a probability c that the feature will be copied correctly, and a probability $1 - c$ that a random feature value will be drawn from the environmental distribution of feature values represented in the system. This draw of a random feature value means that even though the correct feature value may not be stored, the feature value stored will—in some sense—look like a value drawn from the environment.

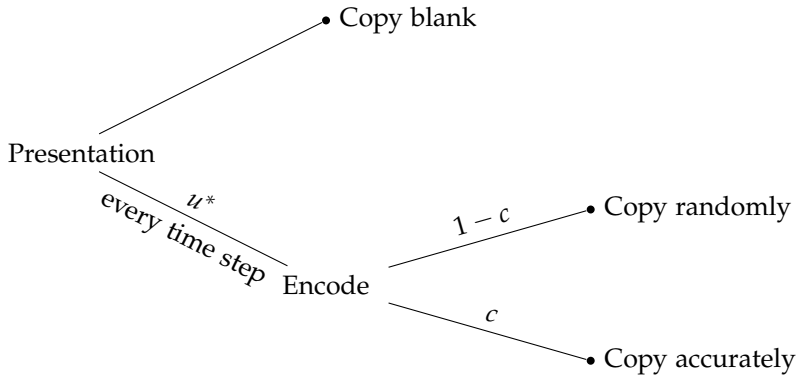

Figure 16: Encoding in REM is in two steps: first, there is the possibility that nothing is encoded at all; second, if something is encoded, it might be encoded incorrectly. Retrieval must take the second two options into account.

In sum, each image is represented as a series of feature values. These feature values have a distribution in the environment. Encoding of items happens over a number of discrete time steps, where there is some probability that an item is encoded. If an item is encoded, there is some probability that it is copied correctly. If the item is not copied correctly, it is replaced with a value drawn from the distribution that generates the feature values in the environment (or an approximation of this distribution known to the system).

Retrieval and comparison

THE PRINCIPAL JUDGEMENT THAT THE MODEL MUST MAKE UPON PRESENTATION OF A PROBE is whether the probe has been randomly drawn from the environment or whether the probe is stored

in memory. This judgement is made through computing a likelihood ratio of the two possibilities against each of the items stored in memory and then averaging the likelihood ratios to obtain an odds ratio. The likelihood ratio is of the form

$$\frac{\text{Likelihood of stored match}}{\text{Likelihood of chance encounter}}.$$

For the purposes of discussion, we will denote the probe as P, the hypothesis that the probe is novel as N and the hypothesis that the probe has been stored as S. Using this notation, the above likelihood ratio becomes

$$\frac{\Pr(P|S)}{\Pr(P|N)}.$$

Computation of the likelihood ratio occurs for each feature independently.

First, it should be noted that the special blank "o" values are uninformative when comparing a probe to a stored memory item. When comparing a probe image to any stored image with blank values, the features in the probe image corresponding to the blank values in the stored image are ignored.

The baseline hypothesis—the denominator of the likelihood ratio—is that the present probe is not a stored image and has simply been drawn randomly from the environment. For a memory system to be able to compute this denominator, it is assumed that some knowledge or sensitivity to the distribution of features in the environment is available. (For reasons stated above, this is likely not to be a controversial assumption, even if it just means that the distribution of features is approximated.) The likelihood of image occurring randomly is simply the product of each of the probabilities of the feature values occurring. That is, assuming that the features are geometrically distributed and the memory system is aware of the distribution parameter,

$$\Pr(P|N) = \prod_{i=1}^{n} x_i = \prod_{i=1}^{n} (1-g)^{V_i - 1} \cdot g$$

where g is the internal representation of the rate parameter, n is the number of features, and V_i is the feature value in the ith position.

The second hypothesis, that the probe image is the stored image being compared against has two separate case. First, if the a given feature in the probe matches one in the stored image, then it was either stored correctly, with a probability c or stored incorrectly with a probability of $1 - c$, but by chance was the right value. In the case that it was stored incorrectly but it was the right value, this is the probability of the value being obtained from the distribution multiplied by the chance of being incorrectly stored, or

$$(1-c)\left[(1-g)^{V_i} \cdot g\right].$$

All considered, the probability that a feature match between a stored image and a probe is

$$c + (1-c)\left[(1-g)^{V_i} \cdot g\right].$$

If a given feature in the probe does not match, it must have arisen through chance after an incorrect encoding. In that case, the probability is equal to

$$(1-c)\left[(1-g)^{V_i}\cdot g\right].$$

This information allows us to compute the likelihood ratio given a probe item and a stored item,

$$\lambda = \prod_{k\in M}\left[\frac{c+(1-c)\left[(1-g)^{V_k}\cdot g\right]}{(1-g)^{V_k-1}\cdot g}\right]\cdot\prod_{j\in Q}\left[\frac{(1-c)\left[(1-g)^{V_j}\cdot g\right]}{(1-g)^{V_j-1}\cdot g}\right]$$

where M is the set of matching non-zero features and Q is the set of non-matching non-zero features.

Given a probe item, these likelihood ratios can be computed across each stored item to produce an odds ratio. Assuming that all items stored in memory are equally likely to be sampled from memory, then the total posterior odds ratio—assuming an unbiased prior odds ratio—is given by

$$\Phi = \frac{1}{n}\sum_{j=1}^{n}\lambda_j$$

where λ_j is the jth image stored in memory.

The default assumption is that a *Phi* value of over 1 should be accepted as a word stored in memory (the average numerator is larger than the average denominator). Likewise, a *Phi* value of less than 1 should be dismissed as a novel (lure) word (the average denominator is smaller than the average numerator). A *Phi* of 1 corresponds to even odds. Although the criterion in terms of *Phi* can be adjusted in the event of uneven prior odds—assuming, *a priori*, that distractors are more or less probable than stored items—the natural criterion is to base the criterion on an even average likelihood ratio.

Empirical predictions

ALL THE EFFECTS IN RECOGNITION MEMORY DESCRIBED SO FAR HAVE NATURAL INTERPRETATIONS IN REM.

The list length effect The list length effect has the most straightforward interpretation in terms of the REM model. Additional items in the study list provide additional sources of variability in the noise/distractor distribution by allowing some distractors to, by chance, match well with novel (lure) probes.

Strength, the list strength effect Stimulus exposure duration is modified in the REM model by changing the number of time steps to encode in the storage of images. This has a straightforward effect of

providing more discriminable information to allow better differentiation between different stored items.

The null list strength effect simply arises as an artifact of the fact that REM predicts that the likelihood calculation in mixed list conditions is calculated equally for strong and weak items, but with weak items possessing less potentially discriminating information.

z-ROC curves The distributions of the likelihood ratios λ and of the posterior odds ratios Φ tend toward Gaussian distributions by virtue of the central limit theorem, giving linear z-ROC functions. The higher variance of the signal distributions, as noted by Ratcliff et al. (1992), arises in REM as a result of unusual feature combinations being stored in memory and giving large likelihood ratios when these items are compared to their matching stored images.

The word frequency effect The word frequency effect is manipulated by changing the geometric rate parameter g across frequency conditions. An assumption is made that high frequency words are more easily confused between each other, consistent with the theoretical explanation of the word frequency effect in terms of signal detection theory. This also assumes that the set of features that typifies high frequency words are also more likely to be concentrated and common in distribution than the features that typify low frequency words. This produces the required effects in d' to be consistent with empirical effects.

The mirror effect The default criterion odds ratio of 1 is analogous to the unbiased criterion in signal detection theory terms. This means that the noise and signal distributions are equally distant from the point of the criterion. If this likelihood ratio is maintained in judgements where the distance between the distributions in terms of noise changes (i.e., where the d' is smaller), then a mirror effect is naturally produced.

References

Anderson, J. R. and Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, 96(4).

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge University Press, Cambridge, England.

Borges, J. L. (1962). Funes the Memorious. In Sturrock, J., editor, *Ficcones*, pages 83–91. Grove Press.

Bower, G. H. (1967). A multicomponent theory of the memory trace. In *The psychology of learning and motivation: Advances in research and theory*, pages 229–325. Elsevier.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, 58(1):17–22.

Donkin, C. and Nosofsky, R. M. (2012). A power-law model of psychological memory strength in short-term and long-term recognition. *Psychological Science*, 23(6):625–634.

Dunn, J. C. (2004). Remember–know: A matter of confidence. *Psychological Review*, 211(2):524–542.

Ebbinghaus, H. (1885). *Über das Gedächtnis. Untersuchungen zur experimentellen Psychologie.* Duncker & Humblot, Leipzig.

Egan, J. P. (1958). Recognition memory and the operating characteristic (Tech. Note AFCRC-TN-58-51). Technical report, Hearing and Communication Laboratory, Indiana University.

Gardiner, J. M. and Java, R. I. (1991). Forgetting in recognition memory with and without recollective experience. *Memory & Cognition*, 19(6):617–623.

Gillund, G. and Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, 91(1):1–67.

Glanzer, M. and Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition*, 13(1):8–20.

Peterson, W. W., Birdsall, T. G., and Fox, W. C. (1954). The theory of signal detectability. In *Proceedings of IRE Professional Group on Information Theory*.

Ratcliff, R., Clark, S. E., and Shiffrin, R. M. (1990). List-strength effect: I. data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(2):163–178.

Ratcliff, R., Sheu, C.-F., and Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review*, 99(3):518–535.

Rumelhart, D. E. and McClelland, J. L. (1986). *Parallel Distributed Processing, Vol. 1: Foundations*. MIT Press, Cambridge, MA.

Shiffrin, R. M. and Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, 4(2):145–166.

Strong, E. K. (1912). The effect of length of series upon recognition memory. *Psychological Review*, 19(6):447–462.

Swets, J. A., Tanner, W. P., and Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review*, 61(5):301–340.

Wixted, J. T. and Ebbesen, E. B. (1991). On the form of forgetting. *Psychological Science*, 2(6).