Variability in recognition memory scales with mean memory strength: implications for the encoding variability hypothesis

Rory W. Spanton[1] & Christopher J. Berry[1]

[1]School of Psychology, University of Plymouth

Author Note

Correspondence concerning this article should be addressed to Rory W. Spanton, School of Psychology, Faculty of Health, University of Plymouth, PL4 8AA. E-mail: rory.spanton@plymouth.ac.uk

The design and analyses in this article were preregistered at the Open Science Framework and are available alongside all experimental materials, data, and analysis code at https://osf.io/wyhb2/.

Abstract

The unequal variance signal detection (UVSD) model of recognition memory assumes that the variance of memory strength for studied items (σ_o) is typically greater than that of non-studied items. It has been proposed that this *old item variance effect* is caused by many factors that affect memory strength at encoding, which add variable amounts of strength to a baseline amount. However, Spanton and Berry (2020) failed to find evidence for this encoding variability hypothesis and instead found that estimates of σ_o tended to be strongly positively correlated with estimates of mean memory strength (d) across participants, suggesting that σ_o may simply scale with d. The present experiments examined the effects of encoding variability and scaling on old item variance by creating conditions in which mean memory strength and encoding variability was either low or high in 2 × 2 factorial designs. In Experiment 1, overall strength determined estimates of σ_o, with no effect of encoding variability. The same effect of overall strength was found in Experiment 2; there was also a significant effect of encoding variability, although this manipulation also had some effect on d and was therefore partially confounded. We conclude that there is strong evidence that mean memory strength at least partially determines estimates of old item variance, and that satisfactory explanations of the old item variance effect should account for this.

Keywords: Recognition Memory, Memory Strength, Encoding Variability, Strength Scaling, Unequal Variance

Variability in recognition memory scales with mean memory strength: implications for the encoding variability hypothesis

In a recognition memory test, participants judge whether they have previously seen items in a particular context. Inevitably, some of these items are remembered better than others. This can be represented in a signal detection model wherein items at test are associated with a ‘memory strength’ (henceforth ‘strength’) variable. The strength of ‘old’ items (those which have been seen in a study phase) and unstudied ‘new’ items are represented as separate Gaussian distributions along a unidimensional continuum. Because of exposure at study, the mean of the old item distribution is generally greater than that of the new item distribution, reflecting a difference in overall strength between the two item types. The difference between these means (*d*) is therefore a measure of recognition performance. Recognition memory judgements are modelled by comparing the strength value of a given item to static criteria along the strength continuum that correspond to different levels of confidence that an item is either old or new. These may range from high confidence that an item is new nearer to the lower end of the continuum, to high confidence that an item is old towards the higher end of the continuum.

Although both new and old items vary in strength, it is widely accepted that the variance of the old item strength distribution (σ_o) is greater than the variance of the new item distribution (see Rotello, 2017, for a review). The acceptance of this *old item variance effect* is motivated by analyses of the *z*-ROC, a *z*-transformed plot of the probability of correctly judging an old item “old” against the probability that a new item is incorrectly judged “old” at each level of recognition confidence in a given response scale. It is commonly found that *z*-ROCs calculated from recognition confidence data are approximately linear, with slopes less than 1 (Glanzer, Kim, Hilford, & Adams, 1999). Since the value of the *z*-ROC slope is equal to the ratio σ_o / σ_n, a non-unit *z*-ROC slope necessitates making σ_o a free parameter with a value typically greater than σ_n. With this parameterization, the unequal variance signal detection (UVSD) model is defined as having parameters $\theta = \{d, \sigma_o, C_1, C_2, \ldots C_I\}$ where *I* is the highest decision criterion level in terms of strength (Kellen, Klauer, & Bröder, 2013). Therefore, the probability of a 'hit' response (a correct 'old' judgement) at criterion *i* according to the model is

$$P(H) = \Phi\left(\frac{d - C_i}{\sigma_o}\right)$$

where Φ is the cumulative normal distribution function. The probability of a 'false alarm' response (incorrectly judging a new item 'old') at C_i is

$$P(FA) = \Phi(-C_i)$$

Although the UVSD model can account for some commonly observed regularities in the *z*-ROC slope (Egan, 1958; Yonelinas & Parks, 2007), its unequal variance assumption was created purely for the need to account for observed data, and not with a priori psychological assumptions in mind. However, a complementary psychological explanation for the unequal variance assumption was later proposed in the form of the *encoding variability hypothesis* (Jang, Mickes, & Wixted, 2012; Wixted, 2007). According to this theory, the old item variance effect is caused by the presence of a very large number of variables that affect memory strength at encoding. These variables contribute additional strength and variance to memory strength across a set of old items during the study phase, resulting in an increase in σ_o relative to σ_n. Examples of such *encoding variables* could presumably include the level of attention paid to a stimulus, intrinsic stimulus properties, item-participant interaction, and many others. Stated mathematically, old items have some level of baseline strength, $B \sim N(\mu_{\text{baseline}}, \sigma_{\text{baseline}})$, which is equivalent to the new item strength distribution (Jang et al., 2012). In the study phase, B is incremented by an added strength variable $A \sim N(\mu_{\text{added}}, \sigma_{\text{added}})$ during encoding. The addition of baseline and added strength gives the resulting old item distribution in the formula $O = B + A$.

There have been several attempts to test the encoding variability hypothesis and compare its predictions with those of other accounts. Koen and Yonelinas (2010) first attempted this in a method where items at study were presented for either a fixed duration of 2500 ms, or a mixture of 1000 and 4000 ms durations. It was found that the latter variable encoding condition did not change estimates of σ_o. Instead, the contribution of an additional recollection process was solely responsible for changes to the *z*-ROC slope, supposedly constituting evidence against the encoding variability hypothesis in favor of a dual-process model. However, subsequent comments by Starns, Rotello, and Ratcliff (2012) and Jang et al. (2012) clarified that these results had no bearing on the encoding variability hypothesis. This was because Koen and Yonelinas's (2010) method mixed two discrete levels of encoding strength, which would be expected to result in a mixture strength distribution rather than a Gaussian product as the encoding variability hypothesis predicts. However, Koen, Aly, Wang, and Yonelinas (2013) later studied the effects of retrieval manipulations on old item variance, finding that it was possible to induce changes in estimates of σ_o without manipulating encoding variability. Although this finding does not exclude the possibility that encoding variability may still have some role in determining estimates of σ_o, it suggests that it is not the only factor that influences old item variance.

More recently, Spanton and Berry (2020) attempted to test the encoding variability hypothesis by manipulating encoding variables directly during study. To avoid the creation of mixture strength distributions that confounded Koen and Yonelinas (2010), encoding variables were manipulated by adding variance along a continuous scale, rather than by mixing two separate conditions of high or low quality encoding. Across three experiments, attempts to influence σ_o by manipulating three encoding variables (study duration, attention, and word frequency) were unsuccessful; there were no resultant effects on σ_o, although each manipulation was assessed to be weak. Despite this, both *d* and σ_o were found to be significantly greater in the low encoding variability condition in Experiment 2, suggesting again that changes in σ_o may result from factors other than encoding variability. Estimates of *d* and σ_o also showed strong positive correlations in every experiment, indicating that old item variance may scale with mean strength. This was not predicted by the encoding variability hypothesis.

The idea that mean memory strength and variance in memory strength are related is evidenced elsewhere in the recognition memory literature. Early global matching models of memory predicted greater old item variance when discriminability increased (Gillund & Shiffrin, 1984; Hintzman, 1988). Although some previous research concluded that the *z*-ROC slope takes a constant value of approximately 0.8 (Ratcliff, Sheu, & Gronlund, 1992; Ratcliff, McKoon, & Tindall, 1994), it was later found that in many cases, increases in mean strength generally decrease the *z*-ROC slope (Glanzer et al., 1999), meaning that mean strength and old item variance increase with one another in several experimental contexts. The finding that greater strength coincides with greater old item variance has since been observed in other studies (Glanzer & Adams, 1990; Heathcote, 2003; Hirshman & Hostetter, 2000; Koen et al., 2013). More recently, Dopkins, Varner, and Hoyer (2017) found that a semantic priming manipulation increased the memory strength of new items and the variance of their corresponding confidence ratings at test, as well as the *z*-ROC slope. This suggests that a form of strength and item variance scaling could apply more generally to both old and new item types – a distribution with a greater mean tends to have a greater variance. In sum, this is evidence that σ_o scales as a monotonically increasing function of *d* in many experimental settings.

The present study aims to test whether estimates of σ_o are affected by encoding variability or mean memory strength. To achieve this, a successful manipulation of encoding variability during the study phase is needed. Despite previous efforts by Spanton and Berry (2020) to add substantial Gaussian variability to individual encoding variables, the resultant effects upon old item variance were weak. This may be because even without experimental manipulation, there are already a very large number of encoding variables that sum to determine levels of added strength in any condition. Therefore, any further attempts to experimentally manipulate a given encoding variable might have a minimal effect on old item variance because added strength already varies greatly. It could also be possible that the effect of any experimentally manipulated encoding variable is partially counteracted by any number of other encoding variables that occur naturally. For example, if word frequency and strength are negatively related whereas concreteness and strength are positively related, then any amount of added strength that a word may receive for having low word frequency may be balanced by a decrement in strength if that word also happens to have low concreteness. Furthermore, there is likely to be a negative correlation between an item's baseline strength value and the increment of added strength it receives during study, which, in conjunction with the aforementioned factors, makes it difficult to establish a strong experimental manipulation of encoding variability (Spanton & Berry, 2020).

A potential way to address these problems is to manipulate multiple encoding variables simultaneously to achieve a greater combined experimental effect upon old item variance. In doing so, the possibility that manipulated encoding variables may systematically counteract each other can also be addressed by ensuring that particular encoding variables are correlated within a word list. Returning to the example above, word frequency and concreteness would be less likely to counteract one another if their values were negatively correlated, increasing their summated effect upon the variance of recognition confidence judgements. Such a condition could be compared with another wherein values of each encoding variable are constrained to be as low in variance as possible, resulting in low encoding variability. Furthermore, if the mean of each encoding variable is equal across word lists in both high and low encoding variability conditions, the overall memorability of stimuli in each set would be controlled. This control of overall stimulus memorability within an encoding variability manipulation allows for memory strength to be manipulated orthogonally as a separate factor.

We aim to do this in two separate experiments. Both experiments were preregistered on the Open Science Framework (https://osf.io/ty8vz/), with details of our main hypotheses, experimental designs, methods, and analyses being disclosed before data collection for each respective experiment. Deviations from our preregistration were also disclosed.

Experiment 1

In the following experiment, we manipulate both encoding variability and memory strength at two levels each (high, low) in a 2 × 2 factorial design. Strength will be manipulated using a one-back digit judgement task identical to that in the ‘fixed’ condition in Experiment 2 of Spanton and Berry (2020). This task will be present as a simultaneous distraction in low strength condition study phases, and absent in high strength conditions. In high encoding variability conditions, words will be selected in a manner that attempts to ensure that they have high Gaussian variance in terms of four normalized variables previously shown to influence memory strength: word frequency (Glanzer & Bowles, 1976), concreteness (Fliessbach, Weis, Klaver, Elger, & Weber, 2006), age of acquisition (AOA; Cortese, Khanna & Hacker, 2010), and word length (Cortese, McCarty & Schock, 2010). Besides word length, each variable will be inter-correlated to promote maximal encoding variability effects. In contrast, words in low encoding variability conditions will have low variance in terms of the above variables (and a fixed word length), with mean word frequency, concreteness, and age of acquisition scores equal to those in high encoding variability conditions.

It is of note that both the old and new words in each encoding variability condition share the same level of manipulated variability in item characteristics. This was to prevent some words in high encoding variability lists being artefactually more discriminable based on their extreme characteristics. However, this decision still leads to a principled manipulation of encoding variability. Despite σ_o being the ratio of new/old item variance in the UVSD model, the encoding variability hypothesis states that in any situation, old items gain added variability purely by virtue of being studied. As the amount of added variability in strength in the encoding variability hypothesis is partially determined by the characteristics of the old stimuli, we would expect our high encoding variability condition to result in more added old item variance than our low encoding variability condition. We therefore have confidence that our encoding variability manipulation is valid, and that it makes a satisfactory attempt to control the discriminability of old and new items.

After fitting the UVSD model to the data, we expect a main effect of our strength manipulation on *d*, with no main effect of encoding variability on *d*, and no interaction. Given this outcome, if mean memory strength influences old item variance we would expect a main effect of strength on σ_o, with no main effect of encoding variability and no interaction. In contrast, the encoding variability hypothesis predicts a main effect of encoding variability on σ_o, with no effect of strength, nor an interaction.

Methods

Participants

64 participants (12 males, 52 females) with a mean age of 22.30 (*SD* = 8.78) from the University of Plymouth Psychology Participation Pool took part in this experiment. Each participant was a University of Plymouth psychology undergraduate, fluent in English as a first language and not dyslexic. Participants received course credits or £8 cash payment for their participation. We justified our sample size on the basis that it was compatible with our partial counterbalancing design (see Design and Procedure), and that it gave us sufficient power to detect a small-medium effect size (i.e., Cohen's $f(V) = .36$, with $\alpha = .05$ and .80 power in a 2 × 2 within-subjects ANOVA).

Materials

A total of 480 unique words were used as stimuli (60 old and 60 new in each condition). Chosen words appeared in the SUBTLEX-UK word database (Van Heuven, Mandera, Keuleers, & Brysbaert, 2014) and databases from Brysbaert, Warriner, and Kuperman (2014) and Kuperman, Stadthagen-Gonzalez, and Brysbaert (2012). Names, proper nouns, and hyphenated words were excluded from an aggregate of the above databases before sampling. Word frequency scores for these words were taken from the SUBTLEX-UK database (Van-Heuven et al., 2014), concreteness scores were taken from Brysbaert et al. (2014), and AOA scores were taken from Kuperman et al. (2012). In high encoding variability conditions, each set of old or new words (four in total) was selected using an algorithm with the following criteria:

1. Words must be 4-10 characters long.

2. Each set of words must have approximately equal mean word frequency (~3), concreteness (~3), and AOA (~10) scores (see Table 1 for exact values).
3. Concreteness and AOA scores must be strongly negatively correlated with word frequency scores within each word list ($r < -.77$ for concreteness and word frequency scores, $r < -.61$ for AOA and word frequency scores, and $r > .26$ for concreteness and AOA scores).
4. The distribution of word frequency, concreteness and AOA scores must not significantly deviate from a normal within each set, according to an Anderson-Darling test ($p > .05$).

The remaining four sets of old/new words in the low encoding variability condition were sampled with the following criteria:

1. Words must be 7 characters long.
2. Each set of words must have approximately equal mean word frequency, concreteness, and AOA scores (with the same constraints as the high encoding variability condition).
3. Each encoding variable must not be highly correlated. Among the word lists generated, word frequency and concreteness had a maximum negative correlation of $r = -.36$. Word frequency and AOA had a maximum negative correlation of $r = -.11$. Concreteness and AOA had a maximum positive correlation of $r = .03$.
4. Word frequency, concreteness and AOA scores must have low variance. For each word, the formula $\Sigma\ |(\mu_e - e_i)|$ was used to determine the summed difference between the mean of each encoding variable (*e*) across all possible words, and its corresponding value in the i^{th} word. The 240 words with the lowest summed difference scores were then randomly sampled from without replacement to create the low encoding variability word lists.

In low strength conditions, participants heard audio clips of a female computer-generated voice speaking a number between 1 and 9 in each trial; this audio was absent in high strength conditions. The whole experiment was conducted on Lenovo desktop computers running an OpenSesame program (Mathôt, Schreij, & Theeuwes, 2012) which displayed all stimuli, instructions, and logged response data. Stimuli were presented in 40 pt ‘mono’ font.

Table 1.

Mean word frequency, concreteness, and age of acquisition scores in Experiments 1 and 2, with standard deviations in brackets.

		Experiment 1		Experiment 2	
Encoding Variable	Word List	High EV	Low EV	High EV	Low EV
Word Frequency	1	2.94 (0.94)	2.95 (0.31)	2.94 (1.21)	2.99 (0.38)
	2	2.88 (0.92)	2.92 (0.36)	2.95 (1.26)	2.97 (0.37)
	3	2.87 (0.95)	2.97 (0.33)	2.89 (1.22)	2.94 (0.34)
	4	2.89 (0.93)	2.94 (0.32)	2.95 (1.26)	2.90 (0.37)
Concreteness	1	3.10 (0.87)	3.14 (0.44)	3.10 (1.14)	3.07 (0.36)
	2	3.09 (0.87)	3.07 (0.51)	3.12 (1.08)	3.18 (0.49)
	3	3.09 (0.87)	3.17 (0.51)	3.11 (1.21)	3.11 (0.48)
	4	3.05 (0.82)	3.09 (0.41)	3.07 (1.12)	3.14 (0.47)
Age of Acquisition	1	10.30 (2.42)	10.40 (0.56)	10.40 (3.33)	10.30 (0.53)
	2	10.30 (2.31)	10.40 (0.51)	10.40 (2.98)	10.30 (0.49)
	3	10.30 (2.52)	10.40 (0.48)	10.40 (3.40)	10.40 (0.58)
	4	10.40 (2.25)	10.20 (0.56)	10.20 (3.69)	10.30 (0.44)

Note: EV = encoding variability.

Procedure

Participants completed all four experimental conditions in a within-subjects design. The order of conditions, the order of high encoding variability word lists, and the order of low encoding variability word lists were all partially counterbalanced according to a Latin square. This resulted in a 4 × 4 × 4 partial counterbalancing design. All participants gave informed consent using a keypress response.

Before participants began their first low strength condition, they completed practice trials where they responded to auditory distractor digits without having to remember items simultaneously. In these practice trials, a fixation point was presented for 500 ms, followed by an auditory digit and a simultaneous visual prompt to respond to the digit from the previous trial. This prompt appeared in the centre of the screen, lasting 3000 ms (on the first practice trial, participants were prompted to make no response as there was no previous trial). This was followed by a 500 ms inter-trial interval (ITI), during which no information was presented in the centre of the screen. The key “Z = Previous number even, M = Previous number odd” remained static near the bottom of the screen for the duration of the practice trials; participants made responses when prompted using the Z and M keys as instructed. To advance to the following study phase, participants had to make eight consecutive correct responses; if they did not do so after 30 trials, the experimenter would re-explain the task to the participant before they attempted the practice trials again.

In each condition of the main experiment, participants then completed a 60-trial study phase. The low strength study phases shared the same trial level procedure as the practice phase, with the exception that instead of a prompt to respond to the previous number, a randomly selected old word was presented in the centre of the screen. In high strength conditions, participants did not have to complete a simultaneous one-back task. Features associated with this task were therefore not present in these conditions, such as the auditory digits and the response key, although the duration of the fixation, stimulus presentation and ITI remained the same. In all conditions, participants were instructed to try their best to pay consistent attention to each word during study.

In between study and test phases in each condition, participants completed a short retention interval in which they answered basic arithmetic questions. These questions took the form "$A \pm B \pm C = ?$" where A, B, and C were one or two digit positive integers. The correct answer was always a one or two digit integer. Participants completed sequential trials of these questions for 60 seconds, at which point they progressed to the test phase.

In every condition, test phases were identically structured; a fixation point would appear for 500 ms, followed by a randomly selected word that was either old or new in the centre of the screen. This word was presented until the participant made a recognition confidence judgement based on their degree of certainty that the item was old or new. Participants made these responses with 1-6 keys at the top of the keyboard, using the category scale "1 – Sure New, 2 – Probably New, 3 – Guess New, 4 – Guess Old, 5 – Probably Old, 6 – Sure Old". This key, and the prompt "New or Old?" were presented near the bottom of the screen as a static reminder of the response categories throughout each test phase. After each response, a 500 ms ITI (in which no information was displayed in the centre of the screen) was displayed, before the next trial. Participants were instructed to make use of the whole rating scale, and to prioritize the accuracy of their judgements over speed as they completed the task.

Results

All analyses were conducted in the statistical programming language R (Version 3.6.3; R Core Team, 2019), primarily using the *tidyverse* package (Wickham et al., 2019). All Bayes

Factors (Scaled JZS) were reported using the *BayesFactor* package (Morey & Rouder, 2018). The UVSD model was fit to the data using maximum likelihood estimation (Dunn, 2010).

In the following analyses we excluded four participants who predominantly used the "Sure New" and "Sure Old" responses, resulting in large outlying parameter estimates. We did so because these data did not give a meaningful representation of variability in recognition responses, and because we defined this criterion for exclusion in our preregistration. We also analysed the natural logarithmic transformation of σ_o because, with the value of σ_n fixed to equal 1, σ_o is a ratio and would otherwise violate the assumptions of a 2 × 2 ANOVA.

Study Task Performance

The proportions of correct responses made in each 'low strength' study phase condition were compared to check whether the presence of an encoding variability manipulation resulted in any task interference effects. The mean proportion of correct one-back task responses did not differ significantly between the "low strength, high encoding variability" condition (M = .94, SE = .01) and "low strength, low encoding variability" condition (M = .94, SE = .01), t(59) = 0.34, p = .74, 95% CI [-0.02, 0.02], BF = 0.15.

Encoding Variability Manipulation

To confirm whether our manipulation of encoding variability influenced subsequent recognition ratings, multiple regression analyses were conducted within each condition for each participant. Word Frequency, Concreteness, Age of Acquisition and Word Length were specified as predictors of recognition confidence ratings for each old item at test. The proportion of significant regression models (as assessed by the F-statistic) and mean R^2 values for each condition are reported in Table 2.

To compare these R^2 values, we conducted a 2 × 2 ANOVA on R^2 with strength (high, low) and encoding variability (high, low) as factors. There was no main effect of strength on R^2, $F(1, 59) = 0.45$, $p = .51$, $\eta_p^2 = .01$, BF = 0.17. However, there was a significant effect of encoding variability on R^2, $F(1, 63) = 47.32$, $p < .001$, $\eta_p^2 = .46$, BF = 1.59×10^9, and no interaction, $F(1, 59) = 0.33$, $p = .57$, $\eta_p^2 = .01$, BF = 0.23. This indicates that the proportion of variance in the ratings explained by the encoding variables increased because of our encoding variability manipulations and not our strength manipulation. R^2 was on average 5-6 % greater in the high encoding variance conditions than the low ones.

Table 2.

The proportion of significant regression models and mean R^2 values (standard deviations in brackets) for each condition in Experiments 1 and 2.

Experiment	Condition	*P*(significant) Regressions	Mean R^2
Experiment 1			
	High Strength, High EV	.22	.11 (.07)
	High Strength, Low EV	.10	.06 (.04)
	Low Strength, High EV	.23	.12 (.07)
	Low Strength, Low EV	.03	.06 (.04)
Experiment 2			
	High Strength, High EV	.38	.13 (.08)
	High Strength, Low EV	.07	.05 (.04)
	Low Strength, High EV	.22	.12 (.09)
	Low Strength, Low EV	.05	.05 (.04)

Note: EV = encoding variability.

Parameter Estimates

All mean UVSD model parameter estimates for each condition are found in Table 3. To compare the influence of our manipulations upon parameter estimates of mean strength from the UVSD model, we conducted a 2 × 2 ANOVA on *d* with strength (high, low) and encoding variability (high, low) as factors. There was a large main effect of strength manipulations on *d*, $F(1, 59) = 42.60$, $p < .001$, $\eta_p^2 = .42$, BF = 2.62×10^{10}. There was no effect of encoding variability on *d*, $F(1, 59) = 0.55$, $p = .46$, $\eta_p^2 = .01$, BF = 0.16, and no interaction was present, $F(1, 59) = 0.51$, $p = .48$, $\eta_p^2 = .01$, BF = 0.24.

Table 3.

Mean parameter estimates (standard deviations in brackets) output by the UVSD model, per condition, in Experiments 1 and 2.

		Condition			
Experiment	Parameter	High Strength, High EV	High Strength, High EV	Low Strength, High EV	Low Strength, Low EV
1	*d*	1.50 (0.76)	1.42 (0.79)	0.92 (0.58)	0.92 (0.71)
	σ_o	1.49 (0.47)	1.41 (0.41)	1.31 (0.33)	1.32 (0.33)
	C_1	-1.02 (1.69)	-1.19 (1.59)	-1.62 (2.79)	-1.59 (1.81)
	C_2	-0.09 (1.63)	-0.13 (1.10)	-0.24 (0.87)	-0.48 (1.61)
	C_3	0.72 (0.46)	0.56 (0.48)	0.51 (0.43)	0.42 (0.41)
	C_4	1.24 (0.61)	1.07 (0.67)	1.14 (0.93)	1.06 (0.83)
	C_5	1.90 (1.04)	1.77 (1.06)	1.93 (1.06)	1.84 (1.15)
2	*d*	2.23 (3.22)	1.60 (1.69)	0.94 (0.68)	0.86 (0.63)
	σ_o	1.84 (1.95)	1.45 (0.79)	1.69 (2.14)	1.28 (0.30)
	C_1	-0.97 (2.64)	-1.18 (2.65)	-1.67 (4.87)	-1.18 (1.56)
	C_2	0.17 (1.51)	-0.41 (2.72)	-0.83 (4.87)	-0.22 (1.20)
	C_3	0.82 (0.57)	0.57 (0.49)	0.49 (0.43)	0.43 (0.52)
	C_4	1.31 (0.68)	1.07 (0.64)	1.00 (0.46)	0.91 (0.55)
	C_5	2.03 (1.12)	1.80 (0.72)	1.69 (0.50)	1.71 (0.88)

Note: EV = encoding variability.

The ordinal pattern of σ_o across conditions followed that of *d*. Another 2 × 2 ANOVA with strength and encoding variability as factors was conducted with σ_o as the dependent variable. A significant main effect of strength was found, $F(1, 59) = 6.20$, $p = .02$, $\eta_p^2 = .10$, BF = 4.75. Again, there was no effect of encoding variability, $F(1, 59) = 0.66$, $p = .42$, $\eta_p^2 = .01$, BF = 0.19, and no significant interaction, $F(1, 59) = 0.75$, $p = .39$, $\eta_p^2 = .01$, BF = 0.25. This is evidence that estimates of σ_o were determined by mean memory strength, rather than encoding variability.

Table 4.

Best fitting regression models relating d and σ_o in each experiment, with R^2 values.

Experiment	Condition	Best Fitting Model	R^2
Experiment 1			
	High Strength, High EV	$\sigma_o = 0.04 + 0.08(d)$	.23
	High Strength, Low EV	$\sigma_o = 0.02 + 0.08(d)$	.24
	Low Strength, High EV	$\sigma_o = 0.02 + 0.10(d)$	.31
	Low Strength, Low EV	$\sigma_o = 0.04 + 0.07(d)$	.26
Experiment 2			
	High Strength, High EV	$\sigma_o = 0.04 + 0.08(d)$	.56
	High Strength, Low EV	$\sigma_o = -0.02 + 0.09(d)$	.50
	Low Strength, High EV	$\sigma_o = 0.03 + 0.08(d)$	.20
	Low Strength, Low EV	$\sigma_o = 0.05 + 0.05(d)$	.09

Note: EV = encoding variability.

Curve-Fitting Analysis

As an exploratory analysis, we fitted linear and polynomial models to estimates of *d* and σ_o to determine the shape of the function by which σ_o scales with *d*. We evaluated three scaling formulae; one in which scaling is linear ($\sigma_o = y + bd$, where y is the intercept), one with linear and quadratic components ($\sigma_o = y + b_1d + b_2d^2$), and one with linear, quadratic, and cubic components ($\sigma_o = y + b_1d + b_2d^2 + b_3d^3$). In a sequential regression procedure, each model was fit to data, and the difference in the fit of each model was computed sequentially using frequentist and Bayesian ANOVAs. Linear models with intercepts between 0.01 and 0.04 and coefficients between .07 and .10 tended to fit the data best (exact values can be found in Table 4). In all conditions, there was no significant improvement in fit being evident in frequentist ANOVAs from adding quadratic, or quadratic and cubic components (ps > .28). Bayesian ANOVAs also supported this conclusion (BFs < 0.44).

Discussion

We found no evidence that encoding variability influenced estimates of old item variance, σ_o, despite our manipulation of encoding variability having a clear impact on recognition confidence ratings. Instead, overall memory strength determined estimates of σ_o. Moreover, curve fitting analyses showed a positive, linear association between d and σ_o, further providing evidence of an association between strength and old item variance. These results provide clear evidence that a strength scaling trend can explain the old item variance effect over and above any contributions of encoding variability in the present experiment.

Although the effect of the encoding variability manipulation on recognition confidence ratings was assessed to be significant, this effect was still relatively small in magnitude (see Table 2 for R^2 values from each condition). It is therefore possible that, assuming the encoding variability hypothesis is true, our manipulation still might not have translated to differences in σ_o that were detectable. This outcome would be unable to explain the presence of the currently observed strength scaling trend; however, it would mean that the encoding variability hypothesis might also hold under a stronger manipulation. In Experiment 2, we aim to establish such a manipulation by adding even more encoding variability to old items than in Experiment 1.

Experiment 2

Although encoding variability was successfully added to old items in Experiment 1, it is possible that the strength of this manipulation was constrained by the Gaussian distributional assumption by which encoding variables were sampled. This assumption was driven by the specification of the encoding variability hypothesis, which states that added strength is Gaussian (Jang et al., 2012). Although this assumption is plausible, the Lyapunov central limit theorem states that many non-identical independent random variables can still sum to a Gaussian form, provided they satisfy certain mathematical assumptions. In practice, it is hard to verify these assumptions since memory strength is a latent variable, but it is possible that adding non-Gaussian added strengths may result a product that is at least close to a Gaussian distribution.

To this end, Experiment 2 will follow a method similar to Experiment 1, although the encoding variables will be permitted to be non-Gaussian. This will maximize encoding variability even more than in Experiment 1, thereby increasing the chance of a stronger manipulation. If this manipulation is successful, the same predicted outcomes from Experiment 1 apply.

Methods

Participants

64 participants (16 males, 47 females) with a mean age of 22.8 (*SD* = 10.7) from the University of Plymouth Psychology Participation Pool took part in this experiment in exchange for either £8 or course participation points. Each participant spoke English fluently as a first

language, was not dyslexic, and had not participated in Experiment 1. Participants were either University of Plymouth psychology undergraduates, or members of the public from the Plymouth area.

Materials and Procedure

Stimuli were 480 words (60 old and 60 new in each condition). These words were sampled with the same constraints as in the previous experiment, with only the following differences:

1. The requirement for the distributions of word frequency, concreteness, and AOA scores to not significantly deviate from a Gaussian in the high encoding variability lists was removed. Instead, the distributions did not strictly adhere to any preset distributional shape and were only constrained to be roughly symmetrical. This was achieved by scoring each word by a weighted index of word frequency, concreteness, and AOA scores, and grouping words based on their distance from the mean of the index, measured in standard deviations. Words were then randomly sampled in equal quantities from each group, resulting in distributions of each encoding variable that were non-Gaussian and had more variance than in Experiment 1.
2. Due to the sampling method, the correlations between each encoding variable were stronger, despite no threshold correlation values being imposed as generative constraints. The negative correlations between word frequency and concreteness ranged between r = -.89 and r = .92. The negative correlations between word frequency and AOA were between r = -.60 and r = -.77. The positive correlations between concreteness and AOA were between r = .50 and r = .65.

The strength manipulation and other materials were identical to the previous experiment. The procedure was also identical to that of Experiment 1, with the only difference being that new word lists replaced those that were previously used.

Results

We excluded four participants who used the “Sure New” and “Sure Old” responses in the majority of test phase trials. These exclusions were made for the same reasons as those in Experiment 1. We also analysed the natural logarithm of $\sigma_{o,}$ as in the previous experiment.

Study Task Performance

As in Experiment 1, the mean proportion of correct responses in the “Low Strength, High Encoding Variability” condition (M = .92, SE = .01) was not significantly different from that in the “Low Strength, Low Encoding Variability” condition (M = .93, SE = .01), $t(59) = -0.31$, p = .76, 95% CI [-0.04, 0.03], BF = 0.15.

Encoding Variability Manipulation

Our encoding variability manipulation was assessed using the same multiple regression analysis as in Experiment 1. The proportion of significant regression models and mean R^2 values for each condition are reported in Table 2. A 2 × 2 ANOVA on R^2 with strength and encoding variability as factors found no main effect of strength on R^2, $F(1, 59) = 0.79$, $p = .38$, $\eta_p^2 = .01$, BF = 0.17. There was, however, a significant effect of encoding variability on R^2, $F(1, 59) = 45.21$, $p < .001$, $\eta_p^2 = .43$, BF = 5.37×10^{12} and no interaction, $F(1, 59) = 0.61$, $p = .44$, $\eta_p^2 = .01$, BF = 0.24. As in Experiment 1, this indicates that the proportion of variance in recognition confidence ratings accounted for by the predictor variables increased between 7-8% with our encoding variability manipulations, and not our strength manipulation.

Parameter Estimates

Mean parameter estimates for Experiment 2 are presented in Table 3. 2 × 2 ANOVAs were conducted to determine whether our encoding variability or strength manipulations influenced estimates of *d*. There was a significant main effect of strength on *d*, $F(1, 59) = 41.96$, $p < .001$, $\eta_p^2 = .42$, BF = 6.67×10^9. There was a significant main effect of encoding variability, although this was accompanied by an inconclusive Bayes Factor, $F(1, 59) = 9.97$, $p = .003$, $\eta_p^2 = .15$, BF = 1.62. There was not a significant interaction, $F(1, 59) = 3.69$, $p = .06$, $\eta_p^2 = .06$, BF = 0.81. Our strength manipulations were therefore shown to have a main effect on *d*, however there was some weak evidence of an effect of encoding variability as well. This suggests that the non-Gaussian manipulation of encoding variability may have had some unintended effect upon memory strength.

To assess whether encoding variability or overall strength influenced estimates of σ_o, we conducted a 2 × 2 ANOVA. There was a significant main effect of strength on σ_o, $F(1, 59) = 12.00$, $p < .001$, $\eta_p^2 = .17$, BF = 47.03. There was also a significant effect of encoding variability on σ_o, $F(1, 59) = 11.16$, $p = .001$, $\eta_p^2 = .16$, BF = 6.71. There was also no significant interaction, $F(1, 59) = 3.22$, $p = .08$, $\eta_p^2 = .05$, BF = 1.21. In sum, there was strong evidence for both an effect of strength and encoding variability on σ_o.

Curve-Fitting Analyses

We conducted the same curve-fitting analyses as in the previous experiment; results from this analysis are found in Table 4. Linear models fitted best in all conditions, as quadratic and cubic components did not improve model fit (*p*s > .18, BFs < 0.70).

Discussion

Unlike in Experiment 1, there was evidence for main effects of both encoding variability and overall strength on estimates of σ_o in Experiment 2. However, contrary to the aims of our study, our encoding variability manipulation significantly affected estimates of *d*, though the Bayes Factor for this result was inconclusive. It is therefore difficult to judge whether some effect of our encoding variability manipulation on σ_o was the genuine result of increased encoding variability, or the consequence of the manipulation also affecting memory strength. However, as the Bayes Factor for the main effect of encoding variability on σ_o was larger than that for the main effect of encoding variability on *d*, it is likely that encoding variability had some effect on old item variance by itself. What is clearer is that our manipulation of memory strength influenced both estimates of *d* and σ_o, and that this is not explained by the current specification of the encoding variability hypothesis.

It is possible that our non-Gaussian encoding variability manipulation gave rise to the unexpected effects of encoding variability on *d*. If the manipulation created a non-Gaussian distribution of old item strength, this could have resulted in some unintended effects upon mean strength, as well as variability in memory strength. Indeed, our encoding variability manipulation in Experiment 1 did not have unexpected effects on *d* as well as σ_o, despite the only major difference between each experiment being the distributional assumption by which encoding variables were sampled. In any case, it is still more certain that overall memory strength has a substantial effect on estimates of σ_o.

General Discussion

Although it has been speculated that encoding variability causes the old item variance effect (Wixted, 2007), previous research has suggested that it cannot solely account for the UVSD model's unequal variance assumption (Koen et al., 2013; Spanton & Berry, 2020). Our results from Experiment 1 reiterate this conclusion, showing that σ_o tends to be determined by mean strength (*d*) in a linear scaling function, with no main effect of encoding variability. In Experiment 2, there was a main effect of encoding variability on σ_o, though this was partially confounded by a weaker effect of encoding variability on *d*. However, overall strength still had the greatest influence on σ_o in this experiment. Therefore, we find evidence that overall memory strength during encoding determined estimates of old item variance in both experiments. In conjunction with the findings of previous research, this constitutes a requirement for future explanations of the old item variance effect to explicitly account for the link between overall memory strength and variability in memory strength.

Although an effect of encoding variability on σ_o was not present at all in Experiment 1, and was not found without a partially confounding effect on *d* in Experiment 2, one might argue that it would be in other circumstances where different encoding variables account for a greater proportion of variance in confidence ratings at test. For instance, we focused on changing stimulus characteristics in accordance with the encoding variability hypothesis, as these provided a practical means to vary several factors simultaneously and measure the effectiveness of our manipulation. It is possible that manipulating encoding variability through factors that were not stimulus characteristics may have resulted in a stronger encoding variability effect, and thus a stronger effect on σ_o. However, it has been established that it is hard to achieve a strong encoding variability manipulation in any currently conceived case due to the specification of the hypothesis itself and the (likely) negative correlation between baseline and added strength (Spanton & Berry, 2020). Despite our attempts to manipulate many potential encoding variables simultaneously, the proportion of variance in recognition confidence ratings accounted for by these encoding variables was still relatively low (e.g., at best, R^2 was 13% on average, in the high strength, high EV condition of Experiment 2), despite being greater than in previous research (Spanton & Berry, 2020). Furthermore, relaxing the distributional assumption associated with the hypothesis in Experiment 2 yielded no conclusive evidence in its favor and instead was a likely contributor to unexpected effects on memory strength. Therefore, as previously concluded in Spanton and Berry (2020), the encoding variability hypothesis is hard to test and falsify.

It is important to clarify that in any case, we do not dispute that items will be encoded to varying degrees during study. This is almost certainly true. However, we find evidence that the currently accepted formalization of this broader theory ($O = B + A$) is limited in several ways. Firstly, as stated above, it is difficult to falsify or test with an empirical manipulation. Secondly, it does not account for evidence that memory strength alone can affect estimates of σ_o, both when manipulated during encoding (as shown in the present experiments) and during retrieval (Koen et al., 2013). Thirdly, it more broadly ignores any other factors that may affect memory strength during retention and retrieval, as it focuses only upon variability at encoding. A successful formal theory of variability in recognition memory should address these three shortcomings to successfully explain the old item variance effect.

It is possible that a dual process theory of recognition memory could integrate better with a strength scaling trend. In the dual process signal detection theory (DPSD; Yonelinas, 1994) model, the proportion of items recollected (R) influences both overall memory strength and variability in memory strength. The model can therefore account for strength scaling through its recollection process, provided that the model's measures of familiarity remain constant. This is a formalization that directly pertains to psychologically meaningful processes and which does not rule out the contribution of factors during retention and retrieval. However, it should also be considered in light of results that do not support a scaling trend. Early work on the topic proposed that although the *z*-ROC slope is commonly less than 1, memory strength manipulations do not affect its supposed value of ~0.8 (Ratcliff, McKoon, & Tindall, 1994; Ratcliff et al., 1992). Although Glanzer et al. (1999) later presented substantial evidence demonstrating that *z*-ROC slopes are not constant, their analyses of previous studies that used stimulus repetition to manipulate memory strength found that the *z*-ROC slope consistently remained unchanged despite significant increases in strength. Glanzer et al. (1999) posited that this variable may fundamentally differ from others for which a scaling function may apply; understanding such a difference in the context of a dual process explanation is likely to provide information about the conditions in which strength scaling occurs, and its psychological cause.

The concept of strength scaling should also be considered alongside research that seeks to distinguish variability in underlying memory representations from variability in decision criterion placement. Much previous work has evaluated signal detection models with added criterion variability components, and such models have potential to add to discussion about the unequal variance assumption (Benjamin, Diaz, & Wee, 2009). Although the addition of equal variability to all decision criteria does not affect the ratio of σ_n / σ_o (Wickelgren, 1968), it has been shown that forms of selective criterion variability can affect the *z*-ROC slope (Cabrera, Lu, & Dosher, 2015). However, such variability may also decrease discriminability depending upon its form, putting such an effect in opposition to a scaling trend. Although this opposition depends upon the specification of the model in question, further development of signal detection models with criterion variability could benefit from accounting for a scaling trend in mean strength and old item variance. By investigating this trend in the context of models with variable criteria, it may be possible to further understand whether strength scaling is a product of a decision process or underlying mnemonic representations. Subject to further verification, such a strength scaling rule could also serve as a utility function in a comparison of both the qualitative and quantitative predictions of different models (Lee et al., 2019). It is therefore worth further investigating the conditions under which strength scaling becomes an empirical regularity to determine whether it can be used as a benchmark by which to evaluate theoretical accounts.

To conclude, we investigated whether changes in the variance of recognition memory strength for old items in the UVSD model were prompted by manipulations of mean strength during study, or encoding variability. In Experiment 1, we found overwhelming evidence that levels of overall memory strength influence old item variance, with no such effect of encoding variability. Although we found evidence for an effect of encoding variability on old item variance in Experiment 2, this was partially confounded, as encoding variability also had some effect on mean memory strength. However, there was again a strong main effect of mean memory strength on estimates of old item variance. We argue that the current specification of the encoding variability hypothesis does not satisfactorily account for these results, and that any future explanation of the old item variance effect should take into consideration the tendency for old item variance to scale with mean memory strength. Going further, investigating the relationship between variability in memory strength and overall memory strength may also aid the understanding of recognition memory more generally.

Acknowledgements

This research was supported by the Economic and Social Research Council (ES/N009916/1). We thank Joel Solomons for his assistance with collecting data for both experiments.

Conflict of Interest Statement

The Authors declare that there is no conflict of interest.

References

Benjamin, A. S., Diaz, M., & Wee, S. (2009). Signal detection with criterion noise: Applications to recognition memory. *Psychological Review*, *116*(1), 84-115. https://doi.org/10.1037/a0014351

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Cabrera, C. A., Lu, Z.-L., & Dosher, B. A. (2015). Separating decision and encoding noise in signal detection tasks. *Psychological Review*, *122*(3), 429- 460. https://doi.org/10.1037/a0039348

Cortese, M. J., Khanna, M. M., & Hacker, S. (2010). Recognition memory for 2,578 monosyllabic words. *Memory*, *18*(6), 595-609. https://doi.org/10.1080/09658211.2010.493892

Cortese, M. J., McCarty, D. P., & Schock, J. (2015). A mega recognition memory study of 2897 disyllabic words. *The Quarterly Journal of Experimental Psychology*, *68*(8), 1489-1501. https://doi.org/10.1080/17470218.2014.945096

Dopkins, S., Varner, K., & Hoyer, D. (2017). Variation in the standard deviation of the lure rating distribution: Implications for estimates of recollection probability. *Psychonomic Bulletin & Review*, *24*(5), 1658–1664. https://doi.org/10.3758/s13423-017-1232-9

Dunn, J. C. (2010). How to fit models of recognition memory data using maximum likelihood. *International Journal of Psychological Research*, *3*(1), 140–149. http://dx.doi.org/10.21500/20112084.859

Egan, J. P. (1958). Recognition memory and the operating characteristic. *USAF Operational Applications Laboratory Technical Note.*

Fliessbach, K., Weis, S., Klaver, P., Elger, C. E., & Weber, B. (2006). The effect of word concreteness on recognition memory. *NeuroImage*, *32*(3), 1413–1421. https://doi.org/10.1016/j.neuroimage.2006.06.007

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, *91*(1), 1-67. https://doi.org/10.1037/0033-295X.91.1.1

Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(1), 5-16. https://doi.org/10.1037/0278-7393.16.1.5

Glanzer, M., & Bowles, N. (1976). Analysis of the word-frequency effect in recognition memory. *Journal of Experimental Psychology: Human Learning and Memory*, *2*(1), 21-31. https://doi.org/10.1037/0278-7393.2.1.21

Glanzer, M., Kim, K., Hilford, A., & Adams, J. K. (1999). Slope of the receiver-operating characteristic in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(2), 500-513. https://doi.org/10.1037/0278-7393.25.2.500

Heathcote, A. (2003). Item recognition memory and the receiver operating characteristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(6), 1210-1230. https://doi.org/10.1037/0278-7393.29.6.1210

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95*(4), 528- 551. https://doi.org/10.1037/0033-295X.95.4.528

Hirshman, E., & Hostetter, M. (2000). Using ROC curves to test models of recognition memory: The relationship between presentation duration and slope. *Memory & Cognition*, *28*(2), 161–166. https://doi.org/10.3758/BF03213795

Jang, Y., Mickes, L., & Wixted, J. T. (2012). Three tests and three corrections: Comment on Koen and Yonelinas (2010). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(2), 513–523. https://doi.org/10.1037/a0025880

Kellen, D., Klauer, K. C., & Bröder, A. (2013). Recognition memory models and binary-response rocs: A comparison by minimum description length. *Psychonomic Bulletin & Review*, *20*(4), 693–719. https://doi.org/10.3758/s13423-013-0407-2

Koen, J., Aly, M., Wang, W.-C., & Yonelinas, A. (2013). Examining the causes of memory strength variability: Recollection, attention failure, or encoding variability? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(6), 1726–1741. https://doi.org/10.1037/a0033671

Koen, J., & Yonelinas, A. (2010). Memory variability is due to the contribution of recollection and familiarity, not to encoding variability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(6), 1536–1542. https://doi.org/10.1037/a0020448

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990. https://doi.org/10.3758/s13428-012-0210-4

Lee, M. D., Criss, A. H., Devezer, B., Donkin, C., Etz, A., Leite, F. P., ... & Vandekerckhove, J. (2019). Robust modeling in cognitive science. *Computational Brain & Behavior*, *2*(3-4), 141-153. https://doi.org/10.1007/s42113-019-00029-y

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324. https://doi.org/10.3758/s13428-011-0168-7

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of Bayes factors for common designs*. Retrieved from https://CRAN.R-project.org/package=BayesFactor

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Ratcliff, R., McKoon, G., & Tindall, M. (1994). Empirical generality of data from recognition memory receiver-operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 763-785. https://doi.org/10.1037/0278-7393.20.4.763

Ratcliff, R., Sheu, C.-F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review*, *99*(3), 518-535. https://doi.org/10.1037/0033-295X.99.3.518

Rotello, C. (2017). Signal detection theories of recognition memory. In J. T. Wixted (Ed.), *Learning and Memory: A Comprehensive Reference* (2nd ed., Vol. 4., pp. 201–225). Elsevier. https://doi.org/10.1016/B978-0-12-809324-5.21044-4

Spanton, R. W., & Berry, C. J. (2020). The unequal variance signal detection model of recognition memory: Investigating the encoding variability hypothesis. *Quarterly Journal of Experimental Psychology.* https://doi.org/10.1177/1747021820906117

Starns, J. J., Rotello, C. M., & Ratcliff, R. (2012). Mixing strong and weak targets provides no evidence against the unequal-variance explanation of zROC slope: A comment on koen and yonelinas (2010). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(3), 793–801. https://doi.org/10.1037/a0027040

Van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521

Wickelgren, W. A. (1968). Unidimensional strength theory and component analysis of noise in absolute and comparative judgments. *Journal of Mathematical Psychology*, *5*(1), 102–122. https://doi.org/10.1016/0022-2496(68)90059-X

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, *114*(1), 152. https://doi.org/10.1037/0033-295X.114.1.152

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1341-1351. https://doi.org/10.1037/0278-7393.20.6.1341

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin*, *133*(5), 800-832. https://doi.org/10.1037/0033-2909.133.5.800