Concreteness and disagreement: Comment on Pollock (2018)

Ian Neath¹ · Aimée M. Surprenant¹

Published online: 6 January 2020
© The Psychonomic Society, Inc. 2020

Abstract

Pollock (*Behavior Research Methods* doi:10.3758/s13428-017-0938-y, Pollock, 2018) points out that most memory experiments using abstract and concrete words have a potential confound: Raters express more disagreement, on average, about the rating for an abstract word than for a concrete word, as evidenced by the larger standard deviation of the rating (SDR). Therefore, past demonstrations of the concreteness effect could be explained by the disagreement hypothesis: Words that engender disagreement (i.e., have a larger SDR) are more difficult to remember than those that engender agreement (i.e., have a smaller SDR). Three experiments test predictions of the disagreement hypothesis. In Experiment 1, concreteness (abstract vs. concrete) and SDR size (small vs. large) were factorially manipulated. A concreteness effect was observed for both SDR sizes, but there was no effect of SDR and there were no interactions involving SDR. In Experiment 2, a concreteness effect was observed despite using abstract words with a small SDR and concrete words with a large SDR, the opposite of what the disagreement hypothesis predicts. Experiment 3 replicated Experiment 2 but with a larger set of stimuli. The results offer no support for the disagreement hypothesis.

Keywords Concreteness effect · Norms · Ratings · Serial recall · Short-term memory

Introduction

Brysbaert, Warriner, and Kuperman (2014) collected concreteness ratings of approximately 40,000 English words. The words were rated on a scale of 1 (abstract) to 5 (concrete), but the norms also provide the standard deviation for each rating (SDR). For some words, there is little variation in the subjects' ratings. For example, *taxi* has a mean concreteness rating of 4.93 with a standard deviation of 0.26, and *spirituality* has a mean concreteness rating of 1.14 with a standard deviation of 0.37. For other words, however, the standard deviation is large, indicating disagreement among subjects. For example, the word *purveyor* has a mean concreteness rating of 3.00 but has a standard deviation of 1.89. In one sense, this makes it an abstract word for some ($3.00 - 1.89 = 1.8$) but a concrete word for others ($3.00 + 1.89 = 4.89$).

Pollock (2018) noted that abstract words typically have a larger SDR than concrete words. Within the Brysbaert et al. norms, words with a rating greater than 3 have a mean SDR of 1.09 compared to 1.19 for words with a rating less than 3. Similarly, words with a rating of 4 or higher have a mean SDR of 0.82 compared to 1.00 for words with a rating of 2 or lower.

Pollock (2018) focused on four studies (Allen & Hulme, 2006; Miller & Roodenrys, 2009; Romani, McAlpine, & Martin, 2008; Walker & Hulme, 1999) which (1) manipulated concreteness, (2) provided their stimuli, (3) used immediate serial recall as the test, and (4) observed a concreteness effect, better recall of concrete words compared to abstract words. In Table 1, we show the mean concreteness rating and SDR for these four studies, broken down further when the study included an additional manipulation; for example, we show the values separately for the high- and low-frequency conditions of Miller and Roodenrys. In addition, we include information from four other studies that meet the same criteria. For all 12 sets of stimuli, a concreteness effect was observed, but for all 12 sets of stimuli, the potential confound noted by Pollock exists: The concrete and abstract words systematically differ in the magnitude of the SDR. In Table 1, the mean SDR for concrete words is 0.72 compared to 1.28 for abstract words. Pollock wrote that

Electronic supplementary material The online version of this article (<https://doi.org/10.3758/s13421-019-00992-8>) contains supplementary material, which is available to authorized users.

✉ Ian Neath
ineath@mun.ca

¹ Department of Psychology, Memorial University of Newfoundland, St. John's, NL A1B 3X9, Canada

Table 1 Mean concreteness rating and standard deviation of the concreteness rating (SDR) for studies using immediate serial recall of abstract and concrete words.

Study	Abstract		Concrete	
	<i>M</i>	<i>SDR</i>	<i>M</i>	<i>SDR</i>
Acheson, Postle, and MacDonald (2010) Similar words	3.33	1.27	4.86	0.44
Acheson et al. (2010) Dissimilar words	3.10	1.24	4.75	0.61
Allen and Hulme (2006)	2.56	1.29	4.75	0.55
Chubala, Surprenant, Neath, and Quinlan (2018)	2.43	1.30	4.56	0.80
Miller and Roodenrys (2009) High frequency	2.35	1.22	4.79	0.47
Miller and Roodenrys (2009) Low frequency	2.40	1.33	4.90	0.34
Neath and Surprenant (2019)	2.77	1.34	4.64	0.69
Romani et al. (2008)	2.77	1.26	4.82	0.48
Tse and Altarriba (2009) Positive words	2.81	1.22	4.50	0.68
Tse and Altarriba (2009) Negative words	2.90	1.34	4.54	0.75
Walker and Hulme (1999) Short words	2.37	1.25	4.55	0.84
Walker and Hulme (1999) Long words	2.26	1.31	4.42	0.85
Mean	2.86	1.28	4.52	0.72

“a skeptic could reasonably argue that these experiments do not actually provide evidence for concreteness effects. The reason is that the comparison being made was meant to be between concrete and abstract items, but the comparison that was actually made was between concrete items, on the one hand, and a group of stimuli about which participants disagree, on the other. It *could* be the case that words that engender disagreement are those that are hard to remember, and that this explains processing differences that have previously been attributed to concreteness/abstractness” (p. 1202, italics in original).

Pollock (2018) then reported three experiments that attempted to produce a concreteness effect when SDR was controlled. The purpose of this comment is first to point out the evidence Pollock offered in favor of the disagreement hypothesis is open to alternative explanations, and second to report three experiments that further compare predictions of the disagreement and concreteness hypotheses.

Each experiment reported by Pollock (2018) used a different set of stimuli and differing methods. In Experiment 1, subjects heard a ten-item list of words spoken by the experimenter and the test was immediate spoken free recall. The abstract and concrete words were equated on a number of dimensions (for details, see Table 2 of Pollock). Most notably, they were equated for SDR (using the Brysbaert et al., 2014, database), and recall did not differ, which is exactly what the disagreement hypothesis predicts. There are two possible issues with this experiment. First, an alternate account is that the

null results were due to differences between the abstract and concrete words. Although the words were equated for frequency according to the SUBTLEX_{UK} norms (van Heuven, Mandera, Keuleers, & Brysbaert, 2014), they differed significantly on a number of other measures of frequency, including CELEX (Medler & Binder, 2005), BNC (van Heuven et al.), and HAL (Balota et al., 2007), with the advantage going to the abstract words. It is possible that the lack of a concreteness effect was due to the higher frequency of the abstract words. A second possible issue is that there were two other conditions, labelled agree and disagree. These both came from the middle of the concreteness rating norms, but differed in the SDR, with the former having a smaller value and the latter a larger value. The disagreement hypothesis predicts better recall of the agree than the disagree words, but no difference was observed.

In Experiment 2, eight pairs of words were shown for 8 s each, followed by a free recall test. (There were also three buffer pairs before and after the critical eight pairs.) As in Experiment 1, the abstract and concrete words were equated for SDR (see Table 6 of Pollock) and no significant effects were reported, as predicted by the disagreement hypothesis. An alternate account, however, is that the abstract word pairs were more semantically related than the concrete word pairs, with this offsetting the usual concreteness advantage. WordNet is an on-line lexical database in which words are organized into synonym sets that represent the underlying lexical concept (Miller, Beckwith, Fellbaum, Gross, & Miller, 1990). Pedersen, Patwardhan, and Michelizzi (2004) calculated a number of measures of similarity between lexical concepts, and the one used here is the number of steps in the shortest path between two words. Thus, low values indicate a

closer semantic relation than high values. The mean path length for abstract word pairs was 7.75 compared to 14.00 for concrete word pairs. It is possible that the lack of a concreteness effect is due to the higher semantic relatedness of the abstract word pairs.

Experiment 3 used six-item lists with typed recall, and as in the previous experiments, the abstract and concrete words did not differ in SDR (see Table 11 of Pollock). Unlike Experiment 1, however, the SDR for both abstract and concrete words were substantially below 1, 0.85 and 0.82, respectively. This time concrete words were better recalled than abstract words. One interpretation is that a true concreteness effect could emerge because now there is sufficient agreement on the ratings. An alternate account, however, is that the results are due to other differences between the abstract and concrete words. For example, the two sets of words differed in frequency, as measured by SUBTLEX_{US} (Brysbaert & New, 2009), CELEX (Medler & Binder, 2005), and HAL (Balota et al., 2007); in age of acquisition (Brysbaert & Biemiller, 2017); and in valence, arousal, and dominance (Warriner, Kuperman, & Brysbaert, 2013). It is therefore not clear what factor or combination of factors caused the difference in memory performance.

Our first point, then, is that the evidence presented in Pollock (2018) for the disagreement hypothesis is open to plausible alternative explanations. On the one hand, many of the results are consistent with the disagreement hypothesis: the null results in Experiments 1 and 2 are due to equating the SDR such that both abstract and concrete words had relatively large values, and that the effect observed in Experiment 3 was due to a reduction in the magnitude of the SDR. However, an alternate interpretation is that differences between the abstract and concrete words on various other measures led to the pattern of results.

The purpose of the three experiments reported below was to directly contrast predictions of the disagreement and concreteness hypotheses. Experiment 1 did this by factorially manipulating both concreteness and the size of the SDR. The disagreement hypothesis predicts better recall for the items with more agreement (smaller SDR) whereas the concreteness hypothesis predicts better recall of the concrete items regardless of the SDR. Experiment 2 did this by comparing recall of abstract words with a small SDR to recall of concrete words with a large SDR. The disagreement hypothesis predicts better recall of abstract than concrete words whereas the concreteness hypothesis predicts the reverse. Experiment 3 did this by replicating Experiment 2 but with a larger stimulus set.

Experiment 1

The purpose of Experiment 1 was to test the disagreement hypothesis by using a 2×2 factorial design with one factor

being whether the word was abstract or concrete and the second factor being whether the word had a small or large SDR. Serial recall was used rather than free recall in order to facilitate comparisons to the studies shown in Table 1. The disagreement hypothesis predicts that performance will depend on the size of the SDR whereas the concreteness hypothesis predicts that performance will depend on whether the word is abstract or concrete.

Subjects Sixty volunteers from ProlificAC were paid £9.00 per hour (prorated) for their participation and were randomly assigned to one of two groups. For the experiments reported here, the following inclusion criteria were used: (1) native speaker of English; (2) approval rating of at least 90% on prior submissions at ProlificAC; and (3) age between 19 and 39 years. The mean age was 30.12 years ($SD = 5.18$, range 19–39) and 30 subjects self-identified as female, 28 self-identified as male, and two self-identified as other. The sample size was set at 30 subjects per group based on the open set condition of Experiment 1 of Neath and Surprenant (2019).

Design The design was a 2 word type (abstract vs. concrete) $\times$ 2 SDR size (small vs. large) $\times$ 6 serial positions mixed factorial. Word type and position were manipulated within subjects and SDR size was manipulated between subjects.

Stimuli Two sets of 38 abstract and 38 concrete words were created, one with a small SDR and the other with a large SDR. The small SDR set had a value comparable to the mean of concrete words in Table 1 whereas the large SDR set had a value comparable to the mean of abstract words in Table 1. Our attempt to equate the words in each set on multiple dimensions was mostly successful (see [Supplementary Materials](#) for details). For the small SDR words, the only significant difference between abstract and concrete words was mean SDR. For the large SDR words, there were two significant differences between the abstract and concrete words. The first was for CELEX frequency (the abstract words were higher, although they were equated on four other measures of frequency) and the second was unconstrained bigram frequency (again the abstract words were higher, although they were equated on constrained bigram frequency). To the extent that these differences might affect the results, it should favor the disagreement hypothesis rather than the concreteness hypothesis by raising performance of the abstract words relative to the concrete words.

Procedure All subjects read an informed consent form and indicated their consent by clicking on a button. They were given instructions about the task both in the consent form and again prior to the first trial. The subjects used a mouse to click on a “Start next trial button.” One second after the fixation point disappeared, six words were shown one at a

time for 1 s each in uppercase letters in the center of the screen. After the final word had been shown, the subjects saw a message that asked them to type in the words they had just seen in strict serial order. They had been informed that they needed to enter the first word first, the second word second, and so on. If they were unsure of a response they were encouraged to guess or else click on a button labelled “Skip.” The subjects initiated the next trial when ready, but this could be done only after six responses had been made.

There were 36 trials, half with concrete and half with abstract words; the order of these trials was randomly determined for each subject. The 18 abstract (and 18 concrete) lists can be thought of as three sets of six trials. Each set of six trials contained six words randomly drawn without replacement from the appropriate pool. After six trials of a given kind, the pool was then replenished. For all subjects, then, each word was seen three times, once in the first six lists for that condition; once in the second six lists for that condition; and once in the last six lists for that condition.

Results and discussion

For all experiments, the data were analyzed using both frequentist and Bayesian techniques using JASP (JASP Team, 2018). For the former, both a p value and an effect size are reported, and for the latter, a Bayes factor (BF) is reported. BF_{10} between 3 and 20 indicates positive evidence for the alternate hypothesis (and therefore evidence against the null hypothesis); BF_{10} between 20 and 150 indicates strong evidence, and BF_{10} greater than 150 indicates very strong evidence (Kass & Raftery, 1995). Similarly, BF_{01} indicates evidence in favor of the null hypothesis. Non-integer degrees of freedom indicate the Greenhouse-Geisser correction was applied because the assumption of sphericity was violated.

The top row of Fig. 1 shows the proportion of words correctly recalled in order as a function of serial position, with the upper left panel showing the data for the large SDR group, and the upper right panel showing the data for the small SDR group. As is apparent in the figure, the size of the SDR did not affect performance whereas concreteness did. This conclusion is supported by the results of a 2 word type (abstract vs. concrete) $\times$ 2 SDR size (small vs. large) $\times$ 6 serial positions mixed factorial ANOVA on the proportion of words correctly recalled in order.

The main effect of word type was significant, $F(1,58) = 40.753$, $MSE = 0.0150$, $\eta_p^2 = 0.413$, $p < 0.001$, $BF_{10} = 11.407$, with better recall of concrete ($M = 0.569$; $SD = 0.185$) than abstract ($M = 0.511$; $SD = 0.173$) words. The main effect of SDR size was not significant, $F(1,58) = 0.018$, $MSE = 0.378$, $\eta_p^2 = 0.000$, $p = 0.893$, $BF_{01} = 4.95$, with equivalent recall in the small ($M = 0.543$; $SD = 0.157$) and large ($M =$

0.537 ; $SD = 0.196$) conditions. There was the usual significant effect of position, $F(2.92, 169.36) = 151.527$, $MSE = 0.034$, $\eta_p^2 = 0.723$, $p < 0.001$, $BF_{10} = 4.512 \times 10^{140}$, revealing typical serial position functions for this type of test.

Importantly, the word type $\times$ SDR size interaction was not significant, $F(1,58) = 0.006$, $MSE = 0.015$, $\eta_p^2 = 0.000$, $p = 0.946$, $BF_{01} = 8.770$. This is consistent with the idea that the concreteness effect does not change as a function of the size of the SDR.

Of the remaining interactions, only one was significant, word type $\times$ position, $F(4.26, 247.23) = 2.756$, $MSE = 0.011$, $\eta_p^2 = 0.045$, $p = 0.026$, although the Bayesian analysis contradicted this, $BF_{01} = 25.161$. To the extent this is a real effect, it reflects a smaller difference between abstract and concrete words at the initial and end positions, and a larger difference in mid-list positions. For the SDR size $\times$ position interaction, $F(2.92, 169.36) = 0.157$, $MSE = 0.058$, $\eta_p^2 = 0.003$, $p = 0.921$, $BF_{01} = 176.03$, and for the three-way interaction, $F(4.26, 247.23) = 1.177$, $MSE = 0.058$, $\eta_p^2 = 0.020$, $p = 0.322$, $BF_{01} = 1.34 \times 10^6$.

The results of Experiment 1 are clear: There is a concreteness effect for both small and large SDR stimuli, and the size of the SDR had no observed effect on performance. This supports the concreteness hypothesis and is inconsistent with the disagreement hypothesis.

One might be concerned that in the small SDR condition, the abstract and concrete words differed significantly in SDR.¹ We do not think this is a concern for four reasons. First, the magnitude of the difference is far smaller than the problem noted by Pollock (2018). The mean SDR is 0.869 for the abstract versus 0.664 for the concrete. In contrast, the typical values in the literature as shown in Table 1 are 1.28 and 0.72, respectively. Second, both mean SDRs are below 1.0, following Pollock’s own recommendation. Third, not only is the mean below 1.0, all individual items have an SDR of 0.99 or less. Finally, the stimuli in the large SDR condition did not differ in SDR and a concreteness effect was still observed.

Experiment 2

Comparing the concrete words with a large SDR (upper left panel of Fig. 1) to the abstract words with a small SDR (upper right panel of Fig. 1) suggests a concreteness effect was observed when the magnitude of the SDR for concrete and abstract words was the *reverse* of what is typically the case, as shown in Table 1. However, this is a between-subject comparison. The purpose of Experiment 2 was to replicate this result, but with two changes. First, the design was changed to a within-subject comparison to mitigate any concerns about

¹ We thank an anonymous reviewer for raising this issue.

Fig. 1 Proportion of words correctly recalled in order as a function of condition and position. The upper left panel shows concrete and abstract words with large standard deviation of the rating (SDR) and the upper right panel shows concrete and abstract words with small SDR, both from

Experiment 1; the lower left panel shows large SDR concrete words and small SDR abstract words from Experiment 2; and the lower right panel shows large SDR concrete words and small SDR abstract words from Experiment 3. Error bars show the standard error of the mean

whether the two groups of subjects were equivalent. Second, a new set of stimuli was used to rule out any stimulus set-specific problems. For this new set of stimuli, the concrete and abstract words differed both in concreteness (1.16 for abstract words compared to 4.21 for concrete) and in SDR (0.79 for the abstract words compared to 1.15 for the concrete). The disagreement hypothesis predicts better recall of abstract than concrete words whereas the concreteness hypothesis predicts better recall of concrete than abstract words.

Subjects Thirty different volunteers from ProlificAC were paid £9 per hour (prorated) for their participation. The mean

age was 30.07 years ($SD = 5.29$, range 19–39 years) and 15 subjects self-identified as female, and 15 self-identified as male.

Design There were two within-subjects factors, word type (abstract vs. concrete) and serial position.

Stimuli The stimuli were 42 abstract and 42 concrete words; the former had a small SDR (0.79) and the latter had a large SDR (1.15). The abstract and concrete words were equated on the same dimensions as in Experiment 1; see [Supplementary Materials](#) for details.

Procedure The procedure was the same as in Experiment 1 except that more lists of a given type could be presented before reusing any words.

Results and discussion

The proportion of words correctly recalled in order is shown in the lower left panel of Fig. 1 and looks the same as the data in the upper row. The data were analyzed by a 2 word type $\times$ 6 serial position repeated-measures ANOVA. The main effect of word type was significant, $F(1,29) = 27.536$, $MSE = 0.016$, $\eta_p^2 = 0.487$, $p < 0.001$, $BF_{10} = 3.283$, with better recall of concrete ($M = 0.569$; $SD = 0.185$) than abstract ($M = 0.511$; $SD = 0.173$) words. The main effect of position was significant, $F(2.12, 61.41) = 48.613$, $MSE = 0.118$, $\eta_p^2 = 0.626$, $p < 0.001$, $BF_{10} = 6.314 \times 10^{52}$, as is usual in serial recall tests. The interaction was not significant, $F(3.98, 115.41) = 0.530$, $MSE = 0.011$, $\eta_p^2 = 0.018$, $p = 0.530$, $BF_{01} = 62.22$.

These results replicate those observed in Experiment 1, but do so in a within-subjects design and also with a second set of stimuli. Even though the concrete words had a significantly larger SDR than the abstract words, a concreteness effect was observed. The disagreement hypothesis predicted the opposite: Words with more agreement (in this case, the abstract words), should have been better recalled than words with less agreement (in this case, the concrete words).

Experiment 3

One potential concern is that in the previous two experiments, subjects saw a given word more than once whereas in the experiments reported by Pollock (2018), each word was seen only once. The purpose of Experiment 3 was to replicate Experiment 2 but ensuring that each word was experienced only once. This was done by creating a new set of 66 abstract words with small SDR and 66 concrete words with large SDR. This allowed for 11 lists per subject for each word type. The predictions are identical to those for Experiment 2: The disagreement hypothesis predicts better recall of abstract than concrete words whereas the concreteness hypothesis predicts better recall of the concrete than abstract words.

Subjects Forty different volunteers from ProlificAC were paid £8 per hour (prorated) for their participation. The mean age was 26.50 years ($SD = 5.79$, range 19–36) and 27 subjects self-identified as female, and 13 self-identified as male. The sample size was increased to 40 due to the decrease in the number of lists per subject.

Design The design was identical to Experiment 2.

Stimuli The stimuli were 66 abstract and 66 concrete words; the former had a small SDR (0.87) and the latter had a large SDR (1.17). The abstract and concrete words were equated on the same dimensions as in Experiments 1 and 2; see [Supplementary Materials](#) for details.

Procedure The procedure was identical to that of Experiment 2, except that no words were repeated and only 11 trials were experienced in each condition.

Results and discussion

The proportion of words correctly recalled in order is shown in the lower right panel of Fig. 1 and shows the same concreteness effect that was observed in Experiment 2. The data were analyzed by a 2 word type $\times$ 6 serial position repeated-measures ANOVA. The main effect of word type was significant, $F(1,39) = 28.950$, $MSE = 0.014$, $\eta_p^2 = 0.426$, $p < 0.001$, $BF_{10} = 4.198$, with better recall of concrete ($M = 0.417$; $SD = 0.164$) than abstract ($M = 0.358$; $SD = 0.161$) words. The main effect of position was significant, $F(2.77, 107.89) = 88.990$, $MSE = 0.059$, $\eta_p^2 = 0.695$, $p < 0.001$, $BF_{10} = 2.415 \times 10^{80}$, as is usual in serial recall tests. The interaction was significant by the frequentist test, $F(4.17, 162.79) = 3.357$, $MSE = 0.013$, $\eta_p^2 = 0.079$, $p = 0.010$, but the Bayesian test indicated positive evidence for null hypothesis, $BF_{01} = 5.39$. To the extent this interaction is real, it is due to a smaller concreteness effect at the final serial positions.

With the exception of the interaction, the results are identical to those of Experiment 2, suggesting that seeing a word multiple times in a study versus just once has no observable effect on performance (see also Neath & Surprenant, 2019).

General discussion

Over the past few years, there has been a huge increase in the availability of various norms. One example is the concreteness norms of Brysbaert et al. (2014). Pollock (2018) noted that in addition to providing researchers with ratings of the key construct, these also provide information about the degree to which the subjects agreed on the rating. Pollock further noted that on average, an abstract word will have less agreement on its rating than a concrete word. One consequence is that past demonstrations of concreteness effects have a potential confound: Concreteness is confounded with agreement. This is an important observation and requires empirical testing.

We argued that the three experiments reported by Pollock (2018) are equivocal on whether the magnitude of the SDR is an important factor and on whether the disagreement hypothesis can account for the results. To remedy this, we reported

three experiments that together showed no observable effect of SDR and instead all showed a concreteness effect. Our conclusion is that the disagreement hypothesis is disconfirmed, at least for concreteness effects. Whether it applies to other dimensions – for example, Warriner et al. (2013) provide both mean rating and SDR for three components of emotion (valence, arousal, and dominance) – awaits appropriate empirical testing.

One key difference between our experiments and those of Pollock (2018) is that the abstract and concrete words used here were equated on more dimensions. For example, the stimuli in Experiment 1 of Pollock differed in frequency, the stimuli in Experiment 2 differed in semantic relatedness, and the stimuli in Experiment 3 differed in frequency, age of acquisition, and three aspects of emotionality (valence, arousal, and dominance). The differences were such that the abstract words were favored in Experiments 1 and 2, where null results were observed, whereas in Experiment 3, the number of differences makes it unclear which one or which combination of differences led to the observed effect. In contrast, the stimuli used in the current experiments did not differ on any of these dimensions. Nonetheless, it is always possible that the abstract and concrete words we used differ on one or more dimensions that we did not control. For this reason, it is critically important that researchers include their stimuli, either in their published papers or as [Supplementary Materials](#). There are many studies of concreteness for which the stimuli are not available, and therefore these papers cannot help in evaluating what factors might be underlying the difference in performance.

Pollock (2018) recommended choosing stimuli such that the SDR is as low as possible, and we fully agree, but for a different reason. Based on our data, we conclude that the disagreement hypothesis does not apply to the concreteness effect, at least for serial recall, but smaller SDRs are associated with more agreement, and more agreement likely indicates a better measure of the underlying construct. We recommend using multiple measures, if possible, for similar reasons: To the extent that the measures converge, it is likely that the stimulus sets better reflect the underlying construct.

References

- Acheson, D. J., Postle, B. R., MacDonald, M. C. (2010). The interaction of concreteness and phonological similarity in verbal working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36, 17–36. doi:<https://doi.org/10.1037/a0017679>
- Allen, R., & Hulme, C. (2006). Speech and language processing mechanisms in verbal serial recall. *Journal of Memory and Language*, 55, 64–88. doi:<https://doi.org/10.1016/j.jml.2006.02.002>
- Balota, D.A., Yap, M.J., Cortese, M.J., Hutchison, K.A., Kessler, B., Loftis, B., Neely, J.H., Nelson, D.L., Simpson, G.B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, 39, 445–459. doi:<https://doi.org/10.3758/BF03193014>
- Brysbaert, M., & Biemiller, A. (2017). Test-based age-of-acquisition norms for 44 thousand English word meanings. *Behavior Research Methods*, 49, 1520–1523. doi:<https://doi.org/10.3758/s13428-016-0811-4>
- Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, Instruments, & Computers*, 41, 977–990. doi:<https://doi.org/10.3758/BRM.41.4.977>
- Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46, 904–911. doi:<https://doi.org/10.3758/s13428-013-0403-5>
- Chubala, C., Surprenant, A. M., Neath, I., & Quinlan, P. T. (2018). Does dynamic visual noise eliminate the concreteness effect in working memory? *Journal of Memory and Language*, 102, 97–114. doi:<https://doi.org/10.1016/j.jml.2018.05.009>
- JASP Team (2018). JASP (Version 0.9.2) [Computer software] <https://jasp-stats.org>
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773–795. doi:<https://doi.org/10.1080/01621459.1995.10476572>
- Medler, D. A., & Binder, J. R. (2005). MCWord: An on-line orthographic database of the English language. Madison, WI: Medical College of Wisconsin, Language Imaging Laboratory. Retrieved from www.neuro.mcw.edu/mcword/
- Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An online lexical database. *International Journal of Lexicography*, 3, 235–244. doi:<https://doi.org/10.1093/ijl/3.4.235>
- Miller, L. M., & Roodenrys, S. (2009). The interaction of word frequency and concreteness in immediate serial recall. *Memory & Cognition*, 37, 850–865. doi:<https://doi.org/10.3758/MC.37.6.850>
- Neath, I., & Surprenant, A. M. (2019). Set size and long-term memory/lexical effects in immediate serial recall: Testing the impurity principle. *Memory & Cognition*, 47, 455–472. doi:<https://doi.org/10.3758/s13421-018-0883-8>
- Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). WordNet::Similarity - Measuring the relatedness of concepts. In S. Dumais, D. Marcu, & S. Roukos (Eds.), *Proceedings of the Fifth Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-2004)*, p. 38–41, Boston. Retrieved from www.aclweb.org/anthology/N04-3012
- Pollock, L. (2018). Statistical and methodological problems with concreteness and other semantic variables: A list memory experiment case study. *Behavior Research Methods* doi:<https://doi.org/10.3758/s13428-017-0938-y>
- Romani, C., McAlpine, S., & Martin, R. C. (2008). Concreteness effects in different tasks: Implications for models of short-term memory. *The Quarterly Journal of Experimental Psychology*, 61, 292–323. doi:<https://doi.org/10.1080/17470210601147747>
- Storkel, H. L. (2004). Methods for minimizing the confounding effects of word length in the analysis of phonotactic probability and neighborhood density. *Journal of Speech, Language, and Hearing Research*, 47, 1454–1468. doi:[https://doi.org/10.1044/1092-4388\(2004\)108](https://doi.org/10.1044/1092-4388(2004)108)
- Tse, C.-S., & Altarriba, J. (2009). The word concreteness effect occurs for positive, but not negative, emotion words in immediate serial recall. *British Journal of Psychology*, 100, 91–109. doi:<https://doi.org/10.1348/000712608X318617>
- van Heuven, W. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for

- British English. *Quarterly Journal of Experimental Psychology*, 67, 1176–1190. doi:<https://doi.org/10.1080/17470218.2013.850521>
- Walker, I., & Hulme, C. (1999). Concrete words are easier to recall than abstract words: Evidence for a semantic contribution to short-term serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25, 1256–1271. doi:<https://doi.org/10.1037/0278-7393.25.5.1256>
- Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45, 1191–1207. doi:<https://doi.org/10.3758/s13428-012-0314-x>
- Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
- This research was supported, in part, by grants from the Natural Sciences and Engineering Research Council of Canada to each author.
- Open Practices Statement: The raw data are available at <https://memory.psych.mun.ca/research/data/j83-data.xlsx> and have also been submitted as Supplementary Materials. The stimuli have also been submitted as Supplementary Materials.