Exploring Trial-level Reliability of Short-term Memory Effects in Immediate Serial Recall

Winston D. Goh (psygohw@nus.edu.sg)
Department of Psychology, National University of Singapore
9 Arts Link, Singapore 117570, Singapore

Abstract

The within-participant reliability of three short-term memory effects – the word length effect, the semantic similarity effect, and the phonological neighbourhood effect – in immediate serial recall was explored using two existing datasets. This was done to address the question of the extent to which individual participants consistently showed the effects across trials, even when the effects were robust at the group-level data. Split-half reliability coefficients were surprisingly low, suggesting that the effects for individual participants were not particularly stable across the experiments. These findings call for more systematic investigations into the extent to which memory effects are reliable across participants.

Keywords: Reliability; immediate serial recall; short-term memory.

Introduction

The short-term memory literature has reported a number of effects that has been well replicated, such as the phonological similarity effect (Conrad, 1964) – lists of dissimilar sounding words are better recalled than lists of similar sound words; the word length effect (Baddeley, Thomson, & Buchanan, 1975) – lists of short words are better recalled than lists of long words; and effects of semantic similarity (Poirier & Saint-Aubin, 1995) – lists of related words are better recalled than lists of unrelated words.

All of these studies have relied on group-level data to infer differences in performance across the various experimental conditions. For example, in the word length effect, an index of how well participants perform in a memory task using short words is obtained by collapsing or averaging scores across all participants in the short word condition. This is then compared statistically with an index of performance for long words by similarly collapsing the scores across participants in the long word condition. A difference in mean performance across the two conditions is then taken as evidence of an effect of the experimental manipulation, word length in this case, on memory performance.

An implication of using mean differences is that theories of memory processes based on observed experimental effects are then built on the assumption of a "prototypical" or "average" human processor. There is nothing inherently wrong with this, given that averaging across participants takes into account variability in the magnitude of the effect across individual participants. Some participants would show a bigger or smaller word length effect than others, but to the extent that aggregated data exhibit this difference across replications of the phenomena, theorists are confident that the observed effects are robust and useful in theorising about the underlying memory mechanisms.

One question, however, that has not been asked as often is the extent to which the same participant would show the same effect within and across experimental sessions. In other words, is the measure of the memory effect stable? This, of course, is the issue of the reliability of a measure or dependent variable.

Reliability refers to the extent to which a measure remains stable and consistent. In the domain of psychometrics and psychological testing, reliability of measurement scales and indices are routinely established. Two common ways in which this is established are using test-retest reliability – the extent to which the measure remains consistent across time – and split-half reliability – the extent to which items within the test are internally consistent. Extrapolating this to the memory domain, this would be asking to what extent participants would show the same magnitude of the effect, e.g. the difference in recall between short and long words. Test-retest reliability can be measured across different sessions of the same experiment, using different sets of stimuli that have been equated on various properties except for the experimental manipulation (this may be more synonymous with the concept of alternate form reliability in the psychometric domain). Split-half reliability can be measured by examining the extent to which performance on one half of a single session is similar to the other half (split according to odd and even numbered trials).

Test-retest reliability of the word length and phonological similarity effects in an immediate memory span task was examined by Logie, Della Sala, Laiacona, Chalmers, and Wynn (1996) in their Experiment 2. Mean span performance was tested on 40 participants one year apart using the same materials for each testing. Surprisingly, the correlations between the effect sizes of both phenomena at the two time points were very poor, ranging between -.31 and .09 (in the psychometric domain, correlations of above .8 are typically desired [Anastasi, 1990, pp. 115]). These results were obtained even though the group-level data replicated both memory effects. It was suggested that the lack of reliability could reflect variable strategies that participants used during the memory span task.

Logie *et al.* (1996) were only able to report test-retest reliabilities as the dependent measure was memory span for the experimentally manipulated word lists, which was not amenable to split-half reliability measures that index the internal consistency of performance within a session, i.e. do people consistently recall short words better than long words across all items and trials? A fixed-length procedure rather than variable lengths that is inherent in the memory span procedure is more amenable to measures indexing the performance between odd and even trials within a session. To the extent that participants switch between strategies across trials, and if this influences the magnitude of the memory effect, one should be able to observe this in the trial-level reliability measures.

While trial-level reliability has not been examined within the memory domain, some researchers in the word recognition field have started asking this question. Stolz, Besner, and Carr (2005) showed that semantic priming effects have surprisingly low within-participant reliabilities. The semantic priming effect refers to the phenomenon where response times in a lexical decision task, where participants make word versus nonword decisions on a target stimulus, is influenced by whether a preceding prime word is related (e.g. *dog*) or unrelated (e.g. *pen*) to the target word (e.g. *cat*). The difference in latencies between the related and unrelated conditions is the priming effect. While this effect is robust and very well documented, it appears that within participants, it is much less stable when assessed for both test-retest and split-half reliability with correlations below .4, sometimes even near zero. The low to moderate reliabilities for semantic priming has recently been replicated by Yap, Hutchinson, and Tan (in press). However, it appears that measures of some of the effects found for isolated word recognition paradigms, e.g. lexical decision without primes or speeded pronunciations to visually presented target words, are more reliable, with correlations above .4 (Yap, Balota, Sibley, & Ratcliff, 2012). These tend to be for the effects of structural features such as number of letters, syllables, and morphemes, whereas network type features such as neighbourhood size (both orthographic and phonological) and semantic neighbourhood density have low to zero correlations.

The implications of these observations for theories of word recognition are beyond the scope of this article. However, the question as to whether there is within-participant reliability for memory phenomena remains unanswered and deserves exploration, given that the field has typically relied on the analysis of group-level data. Do participants show memory effects consistently? This paper reports analyses of two of the present author's past research for which the within-subjects experimental designs facilitate the investigation of split-half reliabilities of the short-term memory effects that were observed.

Analysis 1: The Word Length Effect

Data were taken from Goh and Goh (2006), who examined the interaction between semantic similarity and the word length effect in a 2 (length) x 3 (similarity) x 8 (trial) design. The fully-within subjects design in Experiment 1 was amenable to reliability analyses. Each of the eight trials within one condition comprised 5-word lists where participants had to recall the words in the order they were presented. The two word length conditions comprised short (mostly monosyllabic) and long (mostly trisyllabic) words. The similarity factor comprised three conditions: homogeneous block (where all 8 trials of 5-word lists were from a single category), homogeneous list (where the 5 words in a list were from the same category, but categories changed across lists), and heterogeneous (where all 5 words in a list were from different categories). The six conditions formed by crossing the length and similarity factors were run as blocks in the original study, i.e. all eight trials in one condition were run in one block before moving to the next condition, with counterbalancing done across subjects. The word lists were visually presented.

To assess word-length effects, only the trials from the heterogeneous conditions were examined, since this is the typical manipulation in other word length studies that did not include a semantic similarity independent variable. The data from 50 introductory psychology students who participated for course credit were included.

Figure 1 shows serial recall performance split into the odd and even numbered trials. A within subjects analysis of variance (ANOVA) revealed a main effect of length, *F*(1,49) = 34.56, *MSe* = .02, *p* <.001, showing the standard word length effect where short words are remembered better than long words. Neither the main effect of half nor the interaction were significant, *F*s < 1, indicating that the word length effect generalised across the odd and even halves of the data.

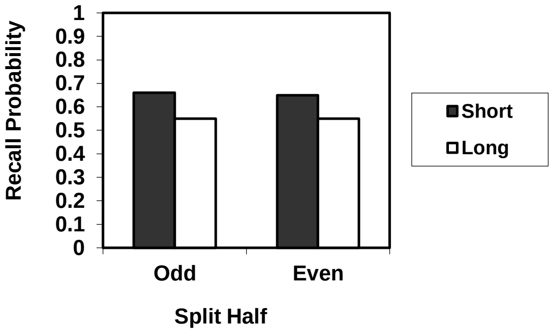

Figure 1. Immediate serial recall performance for lists of short and long words across odd and even numbered trials.

Following the logic reported in the word recognition literature (Stolz *et al.*, 2005; Yap *et al.*, 2012), the difference scores between the short and long words were then computed for each participant for each of the halves. These difference scores were effectively indices of the degree of the word length effect for each participant. The odd and even difference scores were then correlated to obtain the split-half reliability. Figure 2 shows the scatterplot of these difference scores for each participant. If participants exhibited similar magnitudes of the word length effect across the two halves of the experimental session, and hence consistency in the effect, a relatively high correlation should be obtained. The observed correlation was .24, *p* = .098.

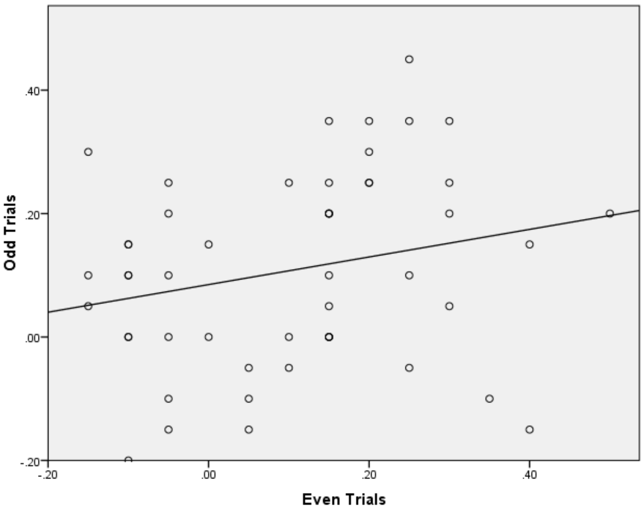

Figure 2. Scatterplot relating the word length effect (short minus long word recall) in odd and even halves for each participant.

Discussion of this finding is deferred to the Discussion section.

Analysis 2: The Semantic Similarity Effect

Data were also taken from Goh and Goh (2006). In line with previous studies on the semantic similarity effect, only the data from the homogeneous list and heterogeneous conditions were examined, since that corresponded to the related versus unrelated conditions in studies looking at effects of semantic similarity. The homogeneous block condition in the original study served a different purpose.

Because short and long words were run as blocks in each of the semantic similarity conditions, the effects of semantic similarity were assessed separately for the short and long words to determine if both showed an advantage of semantic similarity. The data from 50 introductory psychology students who participated for course credit were again included as in Analysis 1.

Figure 3 shows serial recall performance split into the odd and even numbered trials for the short words. A within subjects ANOVA revealed a main effect of semantic similarity, *F*(1,49) = 6.12, *MSe* = .01, *p* <.05, showing the standard semantic similarity effect where related words are remembered better than unrelated words. Neither the main effect of half nor the interaction were significant, *F*s < 1, indicating that the semantic similarity effect for short words generalised across the odd and even halves of the data.

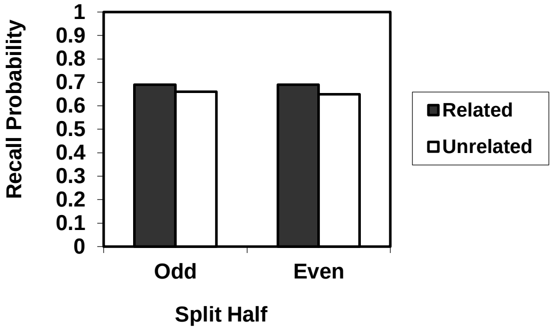

Figure 3. Immediate serial recall performance for lists of related and unrelated short words across odd and even numbered trials.

The difference scores between the homogeneous and heterogeneous conditions were then computed for each participant for each of the halves. These were then correlated to obtain the split-half reliability in order to determine if participants exhibited consistency of the semantic similarity effect across the two halves of the experimental session. Figure 4 shows the scatterplot of these difference scores for each participant. The observed correlation was -.04, *p* = .77.

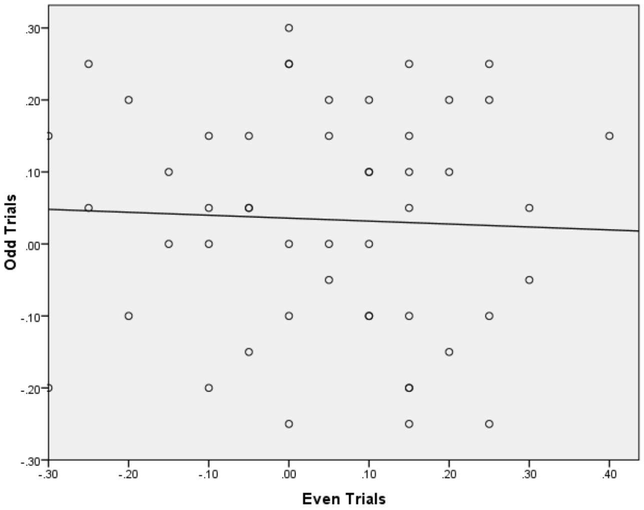

Figure 4. Scatterplot relating the semantic similarity effect (related minus unrelated short word recall) in odd and even halves for each participant.

Figure 5 shows serial recall performance split into the odd and even numbered trials for the long words. A within subjects ANOVA revealed a main effect of semantic similarity, *F*(1,49) = 35.31, *MSe* = .02, *p* <.001, showing the standard semantic similarity effect where related words are remembered better than unrelated words. Neither the main effect of half nor the interaction were significant, *F*s < 1.24, indicating that the semantic similarity effect for long words generalised across the odd and even halves of the data.

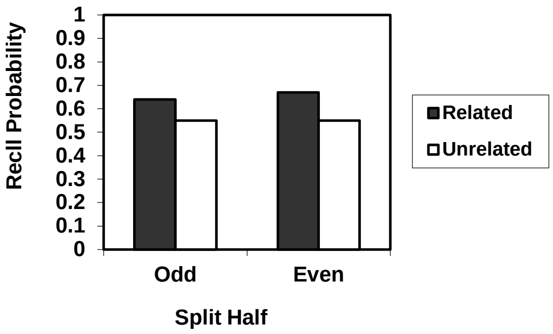

Figure 5. Immediate serial recall performance for lists of related and unrelated long words across odd and even numbered trials.

The difference scores between the homogeneous and heterogeneous conditions were then computed for each participant for each of the halves. These were then correlated to obtain the split-half reliability in order to determine if participants exhibited consistency of the semantic similarity effect across the two halves of the experimental session. Figure 6 shows the scatterplot of these difference scores for each participant. The observed correlation was .26, *p* = .073.

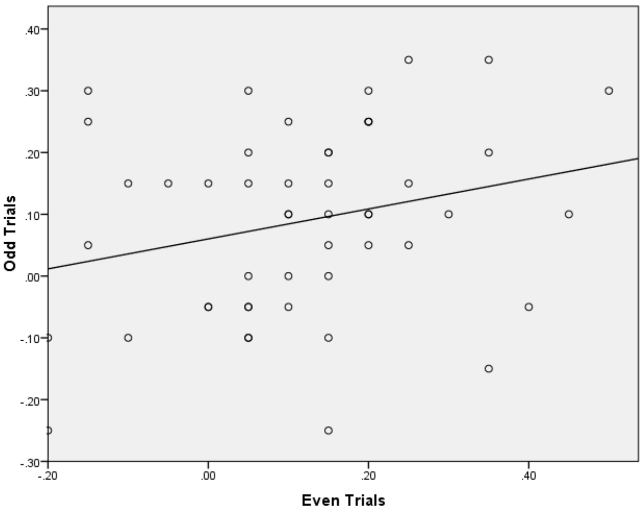

Figure 6. Scatterplot relating the semantic similarity effect (related minus unrelated long word recall) in odd and even halves for each participant.

Analysis 3: The Phonological Neighbourhood Effect

Data were taken from Experiment 2 of Goh and Pisoni (2003), who examined whether immediate serial recall was affected by phonological neighbourhood properties. These include neighbourhood density – the number of words that could be formed from the target word by adding, deleting or substituting a single phoneme – and neighbourhood frequency – the average word frequency of the target word's neighbours. The two word type conditions comprised lexically "easy" words that had low neighbourhood density and low neighbourhood frequency, so that the target words tend to be more distinctive relative to their neighbours; and lexically "hard" words that had high neighbourhood density and high neighbourhood frequency, so that the target words tend to be swamped by their neighbours. There was also a Sampling condition (repeated versus non-repeated) that was run between subjects for a separate purpose. The observed effect was that "easy" words were better recalled than "hard" words in the non-repeated sampling condition.

Only trials from the non-repeated sampling condition, where all words were sampled without replacement, was used. Each of the 11 trials within each word type condition comprised 6-word lists where participants had to recall the words in the order they were presented. The word lists were auditorily presented.

Figure 7 shows serial recall performance split into the odd and even numbered trials. A within subjects ANOVA revealed a main effect of word type, *F*(1,17) = 40.48, *MSe* = .01, *p* <.001, showing that "easy" words are remembered better than "hard" words. Neither the main effect of half nor the interaction were significant, *F*s < 1, indicating that the phonological neighbourhood effect generalised across the odd and even halves of the data.

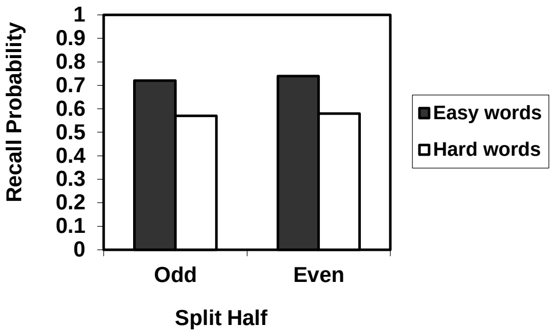

Figure 7. Immediate serial recall performance for lists of lexically easy and hard words across odd and even numbered trials.

The difference scores between the "easy" and "hard" conditions were then computed for each participant for each of the halves. These were then correlated to obtain the split-half reliability in order to determine if participants exhibited consistency of the phonological neighbourhood effect across the two halves of the experimental session. Figure 8 shows the scatterplot of these difference scores for each participant. The observed correlation was .103, *p* = .69.

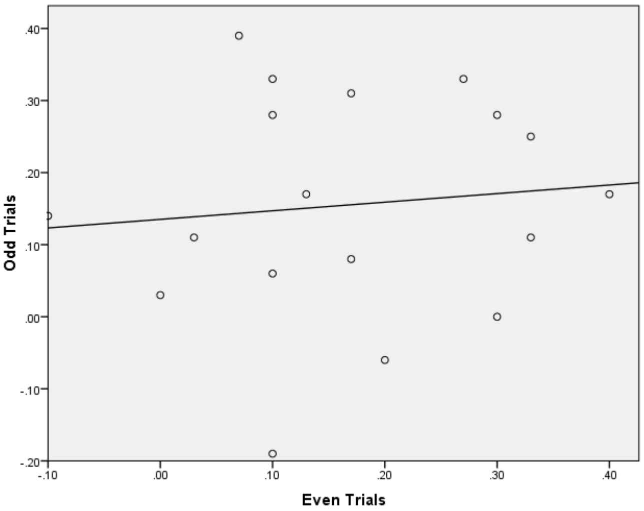

Figure 8. Scatterplot relating the phonological neighbourhood effect (easy minus hard word recall) in odd and even halves for each participant.

Discussion

These exploratory analyses of existing data sets revealed that the observed memory effects were all robust across the split halves for the group-level data – the half factor did not modulate the main effect of interest in all analyses. However, the same cannot be said for within participant reliability. The split-half correlation coefficients obtained were all less than .3, suggesting that the effect for participants was not particularly stable. It should be noted that the magnitude of the coefficients were comparable to some of those observed in the visual word recognition studies cited earlier. For the word length effect in Analysis 1, the coefficient of .24 is higher than Logie *et al*.'s (1996) reported value of -.02 for word length with visual presentation. Although the tasks were different, in that Goh and Goh (2006) used a fixed-length immediate serial recall taks, but Logie *et al*. (1996) used a memory span procedure, the current finding essentially replicated the low reliability observed in the earlier study.

The surprisingly low reliabilities suggest that more systematic investigations ought to be conducted to determine the extent to which this is a phenomenon in these memory effects. The present analyses are limited by the designs of the original studies which were motivated by other research questions. Experiments specifically targeting the reliability issue would include two blocks of different but equated stimuli to facilitate the investigation of test-retest reliability. More trials would also be needed, compared to the 8-11 in the present data sets. Future studies should also be extended to other memory tasks beyond short-term memory that may be amenable to item level analyses, such as recognition paradigms in the long-term memory domain.

The issue of reliability is important as it speaks to the extent to which cognitive effects are stable within participants. It has been suggested that more automatic processes such semantic spreading activation (Posner & Synder, 1976; Stolz *et al*., 2005; Yap *et al*., in press) may show more consistency within participants, but tasks susceptible to attentional control and strategic processes may be more variable. Certainly, the strategies participants may adopt with immediate serial recall tasks (Logie *et al*., 1996) would fall into the latter category. To what extent, therefore, are short-term memory effects influenced by automatic versus controlled processes?

Besides priming effects and other visual word recognition effects reported earlier, it has also been shown that the classic Stroop effect is not reliable in terms of test-retest reliability (Lowe & Rabbit, 1998). The stability of these effects is particularly important if researchers intend to investigate whether individual differences in these effects are associated with other individual differences measures. For example, in the false memory literature, some researchers have investigated the extent to which working memory capacity influences the degree of false recognition (e.g. Peters, Jelicic, Verbeek, & Merckelbach, 2007; Watson, Bunting, Poole, & Conway, 2005). Working memory measures such as forward and backward digit spans, and complex memory spans (e.g. Daneman & Carpenter, 1980; Turner & Engle, 1989) may have established reliabilities, but whether the false recognition indices or measures of other memory effects are also reliable are seldom established. Low reliabilities would inherently limit the extent to which these indices could correlate with other indices, and may lead to erroneous conclusions about the relationship between these effects and other individual differences measures.

In summary, the main goal of the present explorations was to determine if some of the short-term memory effects found in past research using the immediate serial recall task were reliable within participants, as indexed by split-half coefficients. The analyses indicate little to weak reliabilities, although the group-level effects were very robust. These findings call for further, dedicated investigations into the extent to which memory phenomena across other tasks and domains are reliable.

Acknowledgments

This work was supported by Research Grant R-581-000-159-101 to WDG.

References

Anastasi, A. (1990). *Psychological testing* (6^{th} ed.). New York: Macmillan.

Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory.

Journal of Verbal Learning and Verbal Behavior, *14*, 575-589.

Conrad, R. (1964). Acoustic confusions in immediate memory. *British Journal of Psychology*, *55*, 75-84.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*, 450-466

Goh, W. D., & Goh, C. K. (2006). The roles of semantic similarity and proactive interference in the word length effect. *Psychonomic Bulletin & Review*, *13*, 978-984,

Goh, W. D., & Pisoni, D. B. (2003). Effects of lexical competition on immediate memory span for spoken words. *The Quarterly Journal of Experimental Psychology*, *56A*, 929-954.

Logie, R. H., Della Sala, S., Laiacona, M., Chalmers, P., & Wynn, V. (1996). Group aggregates and individual reliability: The case of verbal short-term memory. *Memory & Cognition*, *24*, 305-321.

Lowe, C., & Rabbitt, P. (1998). Test/re-test reliability of the CANTAB and ISPOCD neuropsychological batteries: Theoretical and practical issues. *Neuropsychologia*, *36*, 915-923.

Peters, M. J. V., Jelicic, M., Verbeek, H., & Merckelbach, H. (2007). Poor working memory predicts false memories. *European Journal of Cognitive Psychology*, *19*, 213-232.

Poirier, M., & Saint-Aubin, J. (1995). Memory for related and unrelated words: Further evidence on the influence of semantic factors in immediate serial recall. *Quarterly Journal of Experimental Psychology*, *48A*, 384-404.

Posner, M. I., & Synder, C. R. R. (1975). Attention and cognitive control. In R. Solso (Ed.), *Information processing and cognition: The Loyola symposium* (pp. 55-85). Hillsdale: Erlbaum.

Stolz, J. A., Besner, D., & Carr, T. H. (2005). Implications of measures of reliability for theories of priming: Activity in semantic memory is inherently noisy and uncoordinated. *Visual Cognition*, *12*, 284-336.

Turner, M. L., & Engle, R. W. (1989). Is working memory task dependent? *Journal of Memory and Language*, *28*, 127-154

Watson, J. M., Bunting, M. F., Poole, B. J., & Conway, A. R. (2005). Individual differences in susceptibility to false memory in the Deese-Roediger-McDermott paradigm. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *31*, 76-85.

Yap, M. J., Balota, D. A., Sibley, D. E., & Ratcliff, R. (2012). Individual differences in visual word recognition: Insights from the English Lexicon Project. *Journal of Experimental Psychology: Human Perception and Performance*, *38*, 53-79.

Yap, M. J., Hutchinson, K. A., & Tan, L. C. (in press). Individual differences in semantic priming performance: Insights from the Semantic Priming Project. In M. N. Jones (Ed.). *Big data in cognitive science: From methods to insights*. New York: Psychology Press.