Psychonomic Bulletin & Review (2021) 28:384–408
https://doi.org/10.3758/s13423-020-01815-7

THEORETICAL REVIEW

The detrimental effect of semantic similarity in short-term memory tasks: A meta-regression approach

Sho Ishiguro [1,2] · Satoru Saito [1]

Accepted: 6 September 2020 / Published online: 1 October 2020

Abstract

The literature suggests that semantic similarity has a weak or null effect for immediate serial reconstruction and a facilitative effect for immediate serial recall. These observed semantic similarity effects are inconsistent with the assumptions of short-term memory (STM) models on the detrimental effect of similarity (e.g., confusion) and with observations of a robust detrimental effect of phonological similarity. Our review indicates that the experimental results are likely dependent on the manipulation strength for semantic similarity and that manipulations used in previous studies might have affected semantic assvociation as well as semantic similarity. To address these possible issues, two indices are proposed: (a) strength of manipulation on semantic similarity, gained by quantifying semantic similarity based on Osgood and associates' dimensional view of semantics, and (b) inter-item associative strength, a possible confounding factor. Our review and the results of a meta-regression analysis using these two indices suggest that semantic similarity has a detrimental effect on both serial reconstruction and serial recall, while semantic association, which is correlated with semantic similarity, contributes to an apparent facilitative effect. An effect that is not attributable to similarity or association was also implied. Review on item and order memory further suggests the facilitative effect of semantic association on item memory and the detrimental effect of the semantic similarity on order memory. Based on our findings, we propose a unified explanation of observations of semantic similarity effects for both serial reconstruction and serial recall that is in good accord with STM models.

Keywords Semantic similarity · Short-term memory · Semantic association · Meta-regression

Introduction

The similarity effect on short-term memory (STM) refers to a phenomenon in which the similarity of stimulus properties affects STM performance. STM models generally assume a detrimental effect of similarity (for a review, see Hurlstone, Hitch, & Baddeley, 2014), based on the premise that similarity leads to confusion at retrieval (e.g., the primacy model, Page & Norris, 1998). The SOB (serial order in a box) model (e.g., Farrell, 2006) incorporates the additional premise that similarity decreases encoding weight at encoding. In line with the assumption of the detrimental effect of similarity generally, the detrimental effect of phonological similarity is well replicated using both the serial reconstruction task and the serial recall task (Baddeley, 1966a, 1966b; Baddeley, Lewis, & Vallar, 1984; Conrad, 1964; Poirier & Saint-Aubin, 1996; Watkins, Watkins, & Crowder, 1974; but see also Gupta,Lipinski, & Aktunc, 2005; Nimmo & Roodenrys, 2004). The detrimental effect of auditory similarity is not confined to verbal phonological similarity; non-verbal tonal similarity also has a detrimental effect (Williamson, Baddeley, & Hitch, 2010). Furthermore, the detrimental effect of visual similarity on STM has been demonstrated (Avons & Mason, 1999; Saito, Logie, Morita, & Law, 2008; Smyth, Hay, Hitch, & Horton, 2005; see also Ishiguro & Saito, 2019), suggesting the generality of the detrimental similarity effect on STM.

However, it has also been demonstrated that the detrimental effect of semantic similarity is weak or null in the serial reconstruction task (Baddeley, 1966a, 1966b; Crowder, 1979; Saint-Aubin & Poirier, 1999a; but see also Nelson, Reed, & McEvoy, 1977) and can even be *facilitative* in the serial recall

✉ Sho Ishiguro
ishiguro.sho.grocio@gmail.com

✉ Satoru Saito
saito.satoru.2z@kyoto-u.ac.jp

[1] Graduate School of Education, Kyoto University, Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan

[2] Japan Society for the Promotion of Science, Tokyo, Japan

task (Guérard & Saint-Aubin, 2012; Neale & Tehan, 2007; Poirier & Saint-Aubin, 1995; Saint-Aubin, Ouellette, & Poirier, 2005; Saint-Aubin & Poirier, 1999a; but see also Crowder, 1979).

The results of previous studies on the semantic similarity effect have consequences for STM models, because psychological models rely on observations of effects. For example, a traditional working memory model (Baddeley, 1986; see also Baddeley & Hitch, 1974) assumes a component for storing phonological information, known as the phonological loop, but does not explicitly incorporate a specific component for storing semantic information, because it supposes that verbal STM primarily relies on phonological representations; this is partly because, in early studies on STM, an abundance of data showing robust phonological similarity effects overshadowed the relatively sparse data supporting the semantic similarity effect (for a similar view, see Campoy, Castellà, Provencio, Hitch, & Baddeley, 2015). Although researchers have recently recognized the influence of semantic factors on STM (Bourassa & Besner, 1994; Campoy et al., 2015; Jefferies, Hoffman, Jones, & Lambon Ralph, 2008; Nishiyama, 2014; Walker & Hulme, 1999), another question that memory theories should address is why the semantic similarity effect seems weak or null given that semantic factors affect STM. Moreover, observations of the facilitative effect of semantic similarity lead researchers to suppose there is a process unique to semantics, such as activation of a semantic associative network in an STM model (e.g., Poirier, Saint-Aubin, Mair, Tehan, & Tolan, 2015; Saint-Aubin et al., 2005); therefore, it is theoretically important to consider the results of previous studies on the semantic similarity effect in STM.

In this article, we first overview previous studies and describe existing explanations based on the distinction between item and order memory. Second, we discuss semantic association as a possible confounding factor and suggest that it can be conceptually defined as a pre-experimental associative relationship between the representations of items in long-term memory, and that such semantic associations between items can be operationally defined by the association probability of free association norms. Third, we define and quantify semantic similarity based on Osgood and colleagues' view, which stresses the affective aspect of semantics (e.g., Osgood & Suci, 1955). Fourth, with indices for semantic association and semantic similarity, we review and conduct a retrospective reassessment of previous studies with meta-analytic methods. Finally, we propose a unified explanation for the semantic similarity effect in STM that is in good accord with the assumption of the detrimental effect of similarity. This unified explanation not only provides a means of resolving the inconsistency between the assumption of STM models and observations of semantic similarity effects but also results in novel testable predictions, thereby advancing studies on STM.

Definitions of semantic similarity in previous studies on STM

The semantic similarity effect is typically tested by comparing memory performance on semantically similar words lists with that on semantically dissimilar words lists. Inferior memory performance for similar words lists compared with that for dissimilar words lists supports a detrimental effect of semantic similarity, whereas superior memory performance for similar words lists indicates a facilitative effect. Methods of constructing similar words lists can be classified into three types corresponding to Tse's (2010) three definitions of semantic similarity. First, under the *associative relatedness definition*, words are selected for a similar words list based on their semantically associative relation to each other; for example, a similar words list used in Underwood and Goad (1951) had semantically associated words such as "sunny," "smiling," "festive," and "hearty." Second, by the *categorical relatedness definition*, words on a similar words list are selected based on a taxonomical category (e.g., "apple," "orange," "banana," from the fruit category set). Under the first two methods, dissimilar words lists are typically created by selecting words from different lists of similar words lists or sets (e.g., "apple, piano, gun" from fruit, musical instrument, and weapon categories, respectively, or "amusing, insulting, trivial" from different associatively related lists or sets). This enables the same word to be allocated to both a similar and a dissimilar words list, thereby lessening the influence of individual words' properties, such as frequency and imageability (i.e., it is a form of counterbalancing). The third definition is the *meaning relatedness definition*. Tse (2010) reported that only two studies (Baddeley, 1966a, 1966b), in which criteria for similarity were not specified, met the meaning relatedness definition. Given that any definition of semantic similarity should refer to similarity of meaning, the meaning relatedness definition is likely to entail semantic similarity as subjectively defined by an experimenter, without identifying specific characteristics such as semantic association or taxonomical category. We review previous studies using the serial reconstruction task and/or the serial recall task that meet any of the above-mentioned three definitions of similarity.

Previous studies on semantic similarity in STM

Serial reconstruction Baddeley's classic study (1966a) pioneered research on the semantic similarity effect in STM. In the first experiment, four types of word sequences (namely, semantically similar, semantically dissimilar, phonologically similar, and phonologically dissimilar sequences) were used. In each trial, participants listened to a sequence of five words and wrote down the words in the order in which they were presented. Importantly, the words used in the experiment were presented visually throughout the session. Therefore, the task

was a serial reconstruction task, in which participants were asked for the serial order of stimuli while stimuli were accessible in the test phase. In the serial reconstruction task, stimuli selected at their correct positions are scored (i.e., correct-in-position scoring). Baddeley demonstrated that performance for semantically similar sequences was worse than that for semantically dissimilar sequences; this detrimental effect of semantic similarity was statistically significant but weak compared with the phonological similarity effect, as was also demonstrated by Baddeley (1966a). In another study with a similar procedure (Baddeley, 1966b), a non-significant effect of semantic similarity and a significant effect of phonological similarity were obtained. These studies by Baddeley (1966a, 1966b) have been classified as meeting the meaning relatedness definition of semantic similarity (Tse, 2010).

Later, Crowder (1979) used the serial reconstruction task with associatively and categorically related lists of similar words. Crowder failed to show a significant semantic similarity effect in Experiment 3 but showed a significant detrimental effect of semantic similarity in Experiment 4; however, this was only significant by a one-tailed test. Although the same words were used in both experiments, in the latter experiment, trials of free recall for words were inserted in the serial reconstruction task to prevent strategies such as relying on first-letter cues while orienting participants to the semantic encoding of words, which might have caused differences in the results of the two experiments. Saint-Aubin and Poirier (1999a, Exps. 3 and 4) failed to show a significant semantic similarity effect under the categorical relatedness definition with the serial reconstruction task. Thus, previous studies have suggested that the detrimental effect of semantic similarity in serial reconstruction is at best small. Supposing a small effect size, both results showing statistical significance and those showing non-significance would in principle be reasonable in null hypothesis testing. Some studies, however, even reported a tendency toward semantic similarity advantage (Nelson et al., 1977; Saint-Aubin & Poirier, 1999a, Exp. 3). Taken together, the differences in results across studies imply between-study heterogeneity.

Serial recall The status of the semantic similarity effect in the serial recall task is quite different from that in the serial reconstruction task: the facilitative effect of semantic similarity is well-replicated by a common scoring of correct-in-position, in which items recalled at their correct positions are scored (e.g., Neale & Tehan, 2007; Saint-Aubin et al., 2005; but see also Crowder, 1979). For example, in a study by Poirier and Saint-Aubin (1995), a series of three experiments converged to show a semantic similarity advantage under the categorical relatedness definition. A facilitative effect on STM with a serial recall task was also found using the associative relatedness definition: Tse and associates (Tse, 2009; Tse, Yongna, & Altarriba, 2011) selected 24 themes and constructed lists of associatively related words (e.g., "fast," "molasses," "quick," "snail," "speed," and "turtle") that did not necessarily correspond to a clear taxonomical category, but comprised various words in terms of a word class (such as noun, adjective, and verb). Control lists (i.e., lists of associatively unrelated words) were constructed by drawing words from different themes. Tse and associates demonstrated an advantage for associatively related lists. The facilitative effect of similarity has also been demonstrated with a free recall task (Crowder, 1979), running span task (Kowialiewski & Majerus, 2018), and backward recall task (Guérard & Saint-Aubin, 2012). Thus, in contrast to the weak or null similarity disadvantage for the serial reconstruction task, the similarity advantage is robust for the serial recall and other tasks. Study results suggestive of the similarity advantage are worthy of attention because they are inconsistent with a similarity disadvantage assumed in several STM models that address correct-in-position scoring data (e.g., Brown, Neath, & Chater, 2007; Nairne, 1990; Page & Norris, 1998; for serial reconstruction data, see also Farrell, 2006)[1] and the detrimental effect of phonological similarity on serial recall that is well replicated by correct-in-position scoring (e.g., Page, Madge, Cumming, & Norris, 2007; Poirier & Saint-Aubin, 1996; Watkins et al., 1974).

Item versus order memory distinction and explanations for the semantic similarity effect

Findings on the facilitative effect of semantic similarity for serial recall that seem to contradict the detrimental effect expected by STM models have been addressed by highlighting the distinction between item and order memory. As correct-in-position scoring is thought to reflect both item and order memory, two other scoring methods have been used to investigate item and order memory separately in serial recall (Saint-Aubin & Poirier, 1999b). *Item correct*, also know as the free recall criterion, refers to the number of recalled to-be-remembered items regardless of their positions and is thought to measure item memory. The number of *conditionalized order errors*, in contrast, is the number of to-be-remembered items recalled at their wrong position (i.e., the absolute number of order errors) divided by the number of recalled to-be-remembered items regardless of positions (i.e., the number of item correct), and it is assumed to measure order memory (or its error).

[1] The Scale-Independent Memory, Perception, and LEarning (SIMPLE) model (Brown et al., 2007) is a temporal scale-free model that does not distinguish short-term memory from long-term memory and thus is not a model focusing on STM. The SIMPLE model, however, has been applied to data with STM tasks such as the immediate serial recall task (Brown et al., 2007) and the immediate serial reconstruction task (Surprenant, Neath, & Brown, 2006) and has been used to explain data from STM tasks. In the present article, we aim to investigate data from STM tasks and consider the SIMPLE model as a STM model in the context of STM research.

Separating order memory from item memory enables a fine-grained analysis of the similarity effect. In fact, although the detrimental effect of phonological similarity on serial recall is well replicated in terms of correct-in-position scoring, it has been suggested that phonological similarity is facilitative (or neutral) to item memory but detrimental to order memory (Gupta et al., 2005; Nimmo & Roodenrys, 2004). Similarly, previous studies with the serial recall task have shown that semantic similarity is facilitative to item memory but detrimental (or neutral) to order memory by using both or either of these two scoring methods (Poirier & Saint-Aubin, 1995; Saint-Aubin et al., 2005; Saint-Aubin & Poirier, 1999a, Saint-Aubin & Poirier, 1999b; Tse, 2009; Tse et al., 2011). In addition, as the serial reconstruction task is thought to measure order memory (Saint-Aubin & Poirier, 1999b; Whiteman, Nairne, & Serra, 1994), the detrimental effect of semantic similarity with the serial reconstruction task, although weak, might be seen as the detrimental effect of semantic similarity on order memory.

Based on the distinction between item and order memory, psychological mechanisms for the semantic similarity effect have been proposed by extending a redintegration theory to semantic similarity (e.g., Saint-Aubin et al., 2005; Tse, 2009).[2] A redintegration theory (e.g., Hulme, Maugham, & Brown, 1991; Hulme et al., 1997) supposes a phonological STM process (e.g., Baddeley, 1986) whereby items are encoded and maintained as phonological representations; the process highlights the role of phonological factors in encoding and maintenance. These phonological representations, however, are subject to decay and are not intact at retrieval. The core assumption unique to a redintegration theory is that the redintegration process is necessary for retrieving items to recover degraded phonological representations at retrieval. As evidence supporting a redintegration process separable from the phonological STM process, lexical properties such as lexicality (Hulme et al., 1991) and frequency (Hulme et al., 1997) have been shown to affect STM independently of speech rate (for the separation of a redintegration process from encoding and maintenance processes, see also Schweickert, 1993). The redintegration theory has been extended to support explanations for a semantic similarity effect in at least two ways. First, the process of an *associative network*, with an assumption of a pre-experimental associative relationship between words of associatively or categorically grouped lists, was augmented to a redintegration theory. Tse (2009) implies that the activation of items' representations is enhanced, especially for grouped lists, by the spreading activation of an associative network in long-term memory (Collins & Loftus, 1975), and that these activated parts of an associative network guide a redintegration process. Second, a process of cuing was incorporated into a redintegration theory by assuming that categorically or associatively grouped lists provide *additional retrieval cues* (Poirier & Saint-Aubin, 1995; Saint-Aubin et al., 2005; Saint-Aubin & Poirier 1999a; Tse, 2009). For example, while participants encode a list of "diamond, emerald, and opal" or a list of "bridge, brook, creek," they extract a category or theme label (e.g., "jewel" or "outdoor"). These labels are thought to work as additional retrieval cues in a redintegration process.

According to the extended redintegration theory (e.g., Saint-Aubin et al., 2005; Tse, 2009), semantic similarity is facilitative to item memory due to the effects of an associative network and/or additional retrieval cues that aid the recovery of degraded representations of items. On the other hand, semantic similarity is detrimental to order memory because it leads to overlap between representations, and thus to a form of confusion known as an "interpretation problem": when recalling an item, another item is erroneously recovered and recalled in the position of the targeted one.

Weaknesses of definitions of semantic similarity

We have overviewed previous studies based on definitions of semantic similarity suggested in the literature (Tse, 2010) in order to outline the state of the research on the semantic similarity effect and described existing explanations for the semantic similarity effect in STM in terms of an item versus order memory distinction. In doing so, we have noticed some weaknesses in the studies' definitions of semantic similarity, which are also relevant to existing explanations for the semantic similarity effect based on an associative network and/or additional retrieval cues.

Below, we address possible weaknesses of the associative relatedness definition and the categorical relatedness definition in particular, but not of the meaning relatedness definition, because the assumptions of the first two definitions are intelligible and can be examined, whereas that of the last definition is unspecified and thus intractable.

Associative relatedness definition This definition rests on an assumption that semantically associated words are similar to each other, which is consistent with an explanation of the effects of an associative network: a semantic similarity effect is attributable to an effect by semantic association in an associative network. Nevertheless, the discrepancy between semantic association and semantic similarity, which runs counter to this assumption, has been noted (Hill, Reichart, & Korhonen, 2015; Hutchison, 2003; Tehan, 2010). For example, "cup" is associated with "coffee," but they are not similar in terms of visual information (e.g., solid vs. liquid) or functional information (e.g., container vs.

[2] Tse (2009) uses the term "semantic relatedness" instead of "semantic similarity." However, the list construction methods used in the experiment by Tse (2009) are the methods used for the associative and categorical relatedness definition of semantic similarity by Tse (2010). Considering Tse's terminology, we regard Tse's (2009) extended redintegration theory as an explanation of the semantic relatedness and semantic similarity effects.

beverage). Hill et al. (2015) provided clear evidence showing this discrepancy. They created a database of semantic similarity, called SimLex-999, by asking participants to ignore semantic association and focus on semantic similarity for word pairs. By contrasting similarity values from SimLex-999 with association values (associative strength of word pairs) from the University of South Florida Free Association Database (USF) (Nelson, McEvoy, & Schreiber, 2004), they showed different patterns for similarity and association values; for example, similarity values were higher for synonym pairs than for antonym pairs, while association values were higher for antonym pairs than synonym pairs. Importantly, the comparison of the SimLex-999 and USF databases suggests that participants could distinguish semantic similarity from semantic association, because both databases were based on participants' responses.

Targeting semantic association is critical in memory research because semantic association influences memory performance (Hulme, 2003; Roediger, Watson, McDermott, & Gallo, 2001; Saint-Aubin, Guérard, Chamberland, & Malenfant, 2014; Stuart & Hulme, 2000; Tehan, 2010; Nelson, Bennett, & Leibert, 1997; Tse, 2009; Tse et al., 2011). Roediger et al. (2001) examined data from previous studies using the Deese–Roediger–McDermott (DRM) paradigm, and calculated associative strength between words within a list as *connectivity* by using word association norms. The results suggest that veridical recall of a list increases as the connectivity of a list does. The facilitative effect of inter-item association has also been demonstrated on STM (Hulme, 2003; Saint-Aubin et al., 2014; Stuart & Hulme, 2000; Tehan, 2010; Tse, 2009; Tse et al., 2011). Tehan (2010) demonstrated a beneficial effect of inter-item association in STM, by shortening 15-word lists used in a study employing the DRM paradigm (Stadler, Roediger, & McDermott, 1999) to six-word lists and using these lists in the immediate serial recall task. Furthermore, Tse (2009) demonstrated that the inter-item associative strength of a list (i.e., its connectivity) had a positive linear relationship with performance for immediate serial recall, which extends the findings of Roediger et al. (2001) to the STM domain.

To sum up, it is suggested that (a) semantic similarity can be conceptually separated from semantic association, (b) participants can rate semantic similarity and semantic association separately, and (c) semantic association affects memory performance. Therefore, we suggest that it is important to distinguish semantic similarity from semantic association when identifying and interpreting the semantic similarity effect. A direct comparison of performance for associatively similar versus dissimilar words lists in the previous studies might reflect the effects of both semantic similarity and semantic association.

Categorical relatedness definition The categorical relatedness definition of semantic similarity presumes that words from the same category are similar to each other. It fits well with an explanation that assumes category labels act as additional retrieval cues; the semantic similarity effect is defined by a taxonomical category and as a result, its effect can be explained by the cuing of category labels. However, we have recognized that a category-item relationship is conceptually different from the similarity between items, as in an inter-item relationship. Although category-item relationship (or category membership based on category-item relationship) is an important aspect of semantic structure (Miller, Beckwith, Fellbaum, Gross, & Miller, 1990), we suggest that a taxonomical category does not necessarily correspond to semantic similarity. Moreover, the categorical relatedness definition has led to ambiguity in interpretation and confusion with the concept of semantic association. List construction methods using the categorical relatedness definition are common in studies on the semantic similarity effect (e.g., Saint-Aubin et al., 2005), but this method is also used to manipulate semantic relatedness (e.g., Poirier & Saint-Aubin, 1995). Semantic relatedness may refer to semantic association, because "[t]he semantic relatedness between two items can be defined by their associative strength" (Tse, 2009, p. 874). Thus, results obtained from methods using the categorical relatedness definition have been interpreted as showing evidence for either or both semantic similarity and semantic association, undifferentiatedly; that is, semantic similarity has been confused or identified with semantic association in the literature. In addition to the ambiguity of interpreting the results obtained by list construction methods that are based on the categotical relatedness definition, words on a categorically similar words list (e.g., "cat" and "dog" from the "animal" category) are often both similar and associated with each other. Therefore, we suggest that methods using the categorical relatedness definition affect both semantic similarity and semantic association. This problem is common in the both categorical and associative relatedness definitions.

Semantic association in the current study

As indicated, we argue that semantic similarity has often been confused with semantic association and that the results of previous studies using both associative and categorical relatedness definitions have been affected by semantic association as a confounding factor. Semantic association in this sense can be interpreted as a pre-experimental associative relationship between words and quantified via the associative strength values of free association norms (i.e., connectivity), which allows statistical control of the effect of semantic association on memory performance. Given that the similarity effect has been tested using the difference in performance on similar and dissimilar words lists, we speculate that the *connectivity difference* (CD) between lists of similar and dissimilar words contributes to the apparent advantage of similarity. Therefore, we suppose that a study with a large connectivity difference is

likely to show the apparent facilitative effect of semantic similarity.

Although semantic association can be conceptually and statistically separated from semantic similarity, it remains unclear exactly what semantic similarity is and how it is quantified. In the next section, we address these questions.

Semantic similarity in the current study

Dimensional approach to semantics A feasible approach to studying stimulus properties would be to identify critical dimensions; such a dimensional or elementalist approach has been fruitful in visual STM research, in which dimensions such as location, shape, and color are presumed and the effects of these dimensions are systematically examined (e.g., Logie, Brockmole, & Jaswal, 2011). Furthermore, studies examining phonological similarity using rhyming versus alliterative lists (e.g., Gupta et al., 2005) can be viewed as focusing on dimensions of phoneme overlap, because the position of each phoneme can be regarded as constituting a dimension. In fact, phonological similarity can be operationalized in several ways based on which dimensions are targeted (e.g., last phoneme(s) for rhyming) and its effect depends on targeted dimensions (Gupta et al., 2005; Nimmo & Roodenrys, 2004).

Although they may not be as intuitive as visuospatial dimensions or dimensions of phoneme overlap, *dimensions of semantics* have been also investigated, that is, studies have taken a dimensional approach to semantics (Henley, 1969; Osgood & Suci, 1955; Rips, Shoben, & Smith, 1973). We argue that valence, arousal, and dominance dimensions are particularly important in semantics, based on Osgood and associates' view of semantics (e.g., Osgood & Suci, 1955) and recent findings from computational natural language processing research (e.g., Hollis & Westbury, 2016).

A pioneering work by Osgood and Suci (1955) investigated semantic dimensions by applying factor analysis to data collected by the semantic differential method (Osgood, 1952). Using this method, participants were asked to write down descriptive adjectives for given nouns (e.g., "big" for "house") and these adjectives were collected and used to construct scales (e.g., a graphical 7-point scale of "big to small"). Then, participants rated various words that were different from the original nouns on these scales. Osgood and Suci demonstrated that three factors accounted for approximately 50% of the total variance of the ratings by factor analysis. While these three major factors of semantics were labeled *evaluation*, *activity*, and *potency* in early studies (Miron, 1969; Osgood, 1969; Osgood & Suci, 1955), these labels were replaced with *valence*, *arousal*, and *dominance,* respectively, in later studies, to stress their affective connotations (Bradley & Lang, 1999; Warriner, Kuperman, & Brysbaert, 2013; see also the Pleasure–Arousal–Dominance model, Mehrabian, 1996). It should be noted that the procedure of the original study, which used the semantic differential method (Osgood & Suci, 1955), was not specifically designed to target affective connotations of semantics, but major dimensions of semantics turned out to be related to emotions (Osgood, 1969). In this sense, according to the view of Osgood and associates, dimensions relevant to emotions do not only underlie an emotional part of semantics but also underlie *the entire of semantics*. To sum up, the view of Osgood and associates supports the assertion that semantics can be represented in terms of dimensions and that three particular dimensions – valence, arousal, and dominance – are major semantic dimensions.

Recent computational linguistics models such as Latent Semantic Analysis (LSA) (Landauer & Dumais, 1997) and word2vec (Mikolov, Chen, Corrado, & Dean, 2013) can be seen as modern dimensional approaches to semantics (for another model with a dimensional approach, Hyperspace Analogue to Language (HAL), see Lund & Burgess, 1996). Based on text corpus data, both LSA and word2vec offer a multidimensional space that represents semantic information. The validity of LSA's space has been demonstrated by comparing human data for multiple-choice test items on the Test of English as a Foreign Language (TOEFL) with LSA's results for the same test items; calculation of a cosine value between vectors of a problem word and four alternative words on around 300 dimensions and selection of the alternative word with the highest cosine value provided correct answers as frequently as applicants from non-English-speaking countries to colleges in the USA (Landauer & Dumais, 1997). Landauer and Dumais (1997) noted that the average score of these applicants was acceptable to many universities, suggesting that LSA was able to mirror responses of adequately proficient English speakers. Furthermore, research has demonstrated that simple algebraic operations for vectors in a multidimensional space provided by word2vec can mimic human-like reasoning: the vector of "king" – that of "man" + that of "woman" results in a vector close to the vector of "queen" (for other examples, see Mikolov et al., 2013). These results suggest that such a multidimensional space represents semantic information (i.e., semantic space) and that inspecting the semantic space can provide insights for meaning.

Hollis and Westbury (2016) highlighted the importance of valence, arousal, and dominance dimensions in semantics by investigating a semantic space provided by word2vec. They applied principal component analysis (PCA), which reduces the dimensionality and extracts principal components (PCs) of data, to a 300-dimensional semantic space provided by word2vec and investigated the relationship between the extracted PCs and the semantic variables in word norms (Warriner et al., 2013). The results showed that semantic variables of valence, arousal, and dominance were correlated with the fifth, second, and seventh PCs, respectively (rs = .50, .27, .38, respectively), which suggests that these three dimensions are major dimensions of semantics. If valence,

arousal, and dominance dimensions are indeed major dimensions in semantics, then it is expected that a semantic space will contain these three types of information. For example, the valence value of a word would be estimated based on the valence values of neighboring words in the semantic space. Bestgen and Vincze (2012) demonstrated that three types of values of words rated by human participants were correlated with values estimated based on 30 neighboring words in a semantic space provided by LSA (correlations between human ratings and estimated values were *r*s = .71, .56, .60 for valence, arousal, and dominance, respectively) (for a replication of Bestgen & Vincze's results, see Recchia & Louwerse, 2015).[3]

Although specific dimensions might be critical for a particular domain of concepts (e.g., *body size* and *predacity* dimensions in concepts of animals, Henley, 1969; Rips et al., 1973), we assume that the findings from Osgood and Suci (1955) using words sampled from across domains and from Hollis and Westbury (2016) using a large sample of words (N = 12,344) are representative of semantics across domains, and hence, that valence, arousal, and dominance are common dimensions across domains. We do not argue that other dimensions are irrelevant to semantics, but assume that the three targeted dimensions are major dimensions of semantics as a working hypothesis. In other words, we suggest that valence, arousal, and dominance are primary dimensions that contribute to our understanding of semantic complexity. In addition, this assumption enables a quantitative estimation of semantic similarity.

Our proposition of a semantic space and semantic similarity With the three dimensions of valence, arousal, and dominance, we can create a three-dimensional semantic space based on these three types of values using existing norms (Warriner et al., 2013).[4] By regarding spatial distance as dissimilarity or spatial proximity as similarity (Brown et al., 2007; Lund & Burgess, 1996; Rips et al., 1973), semantic similarity can be viewed as how close words are to each other on the valence-arousal-dominance dimensional space.[5]

Specification of our proposed three-dimensional space To show examples of the spatial representation of semantics, we created two example word lists based on the materials used by Tse et al. (2011) and plotted these words on the three-dimensional space (Fig. 1). An example similar words list is "diamond, emerald, opal, pearl, ruby, sapphire" and an example dissimilar words list is "diamond, aunt, iron, captain, cat, bishop." Visual inspection suggests that the words on the similar words list are placed close to each other while those on the dissimilar words list are placed dispersedly, which fits our assumption that semantic similarity can be viewed as proximity on the valence-arousal-dominance dimensional space.[6] Moreover, this spatial representation offers a potential explanation of what words are similar: similar words are those that are close in valence, arousal, and dominance.

Semantic similarity index based on our proposed three-dimensional space As a formal quantification, we propose *mean distance from the centroid (to each word)* as an index for list dissimilarity (see also Mewhort et al., 2018). This refers to the averaged value of distances between the centroid of words on a list and each word. The mean distance from the centroid would be relatively small for similar words lists because words on these lists are placed closely together, whereas it would be large for dissimilar words lists. As for the two

[3] Findings by Hollis and Westbury (2016) and Bestgen and Vincze (2012) provide clear evidence suggesting that a semantic space provided by a computational model represents semantics. However, such a semantic space is also thought to represent lexical/syntactic information such as co-occurrence and syntactic regularities (Landauer & Dumais, 1997; Mikolov et al., 2013), which suggests that a semantic space by a computational model may reflect information above and beyond semantics. In the analysis, we use semantic similarity indices based on both a semantic space by a computational model (i.e., word2vec) and our proposed semantic space for comparison purposes.

[4] Determining dimensions and estimating similarity based on the dimensions of valence, arousal, and dominance has pragmatic advantages. Suppose that we have k dimensions and n target words. Similarity can be calculated based on kn ratings; thus, the number of necessary ratings is linear with n. In contrast, if we base similarity on similarity ratings for pairs of target words, the number of ratings would be ${}_nC_2$, which has the term of square of n; therefore, the number of necessary ratings increases rapidly as n increases.

[5] Semantics can also be expressed as featural representations (McClelland & Rogers, 2003; McRae, Cree, Seidenberg, & McNorgan, 2005; Vinson & Vigliocco, 2008). For example, "snake" does not have the feature "has legs" while "dog" does. A pattern of several features can represent a concept. Although a featural representation seems different from a dimensional representation, it can also be seen as a point in (featural) space (Vigliocco, Vinson, Lewis, & Garrett, 2004). We suggest that featural representations are a special case of dimensional representations with binary values and, thus, are compatible with dimensional representations.

[6] Although cosine similarity is an oft-used measure in multidimensional models (e.g., Mikolov et al., 2013), we adopted Euclidean distance to represent similarity (see also Lund & Burgess, 1996; Mewhort, Shabahang, & Franklin, 2018) given that distance between two items (i.e., Euclidean distance) is likely to represent similarity more accurately than the angle between the two vectors of two items (i.e., cosine similarity) on our three-dimensional space. This is because the dimensions of the current three-dimensional model have specific meanings (i.e., valence, arousal, and dominance) unlike abstract dimensions provided by computational models. Suppose that an item can be expressed as a vector of three elements respectively representing valence, arousal, and dominance (or a point on the three-dimensional space) and that there are three items, item A [1, 1, 1], item B [1, 2, 1], and item C [9, 9, 9]. Although item A and B differ in arousal values by only one point while items A and C differ by eight points in each value, cosine similarity indicates that item A is more similar to item C (cosine = 1) than to item B (cosine = 0.94). A cosine similarity measure inappropriately shows high similarity between item A (a word with low-valence, low-arousal, and low-dominance values) and item C (a word with high-valence, high-arousal, and high-dominance values). In contrast, the Euclidean distance measure indicates that item A is more similar to item B (distance = 1) than item C (distance = 13.9). Formal calculations of similarity measures in this study are provided in the *Method* section.

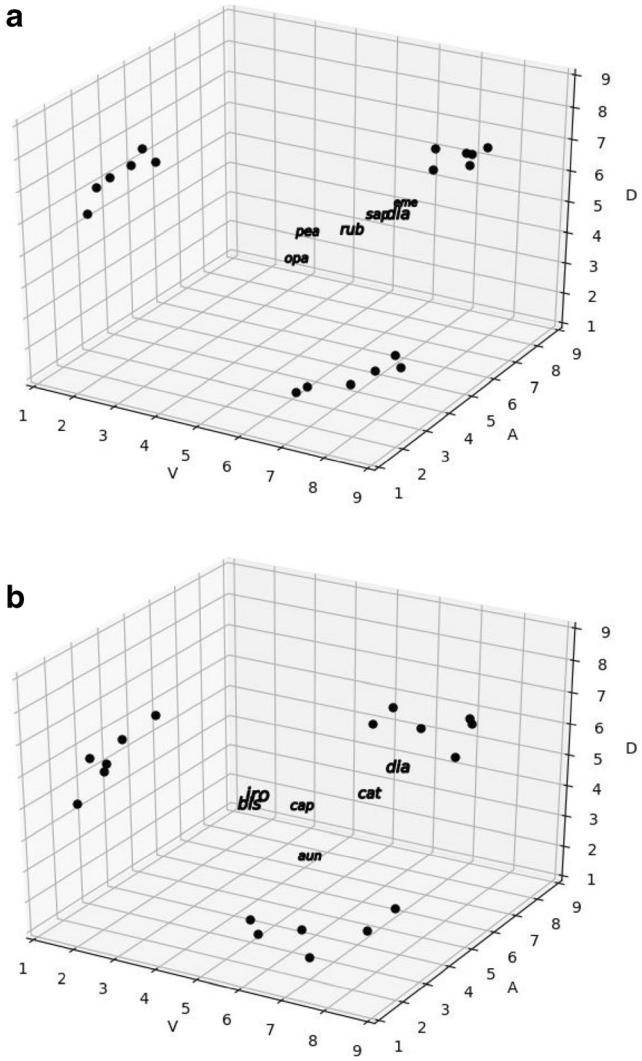

Fig. 1 Representations of words in the valence-arousal-dominance semantic space: (a) an example of a similar words list; (b) an example of a dissimilar words list. The first three letters of each words are depicted. V, A, and D refer to valence, arousal, and dominance respectively

example lists that we created based on Tse et al. (2011), the mean distance from the centroid for the example similar words list is 1.04 and that for the example dissimilar words list is 1.33. We propose that the difference between the mean distance from the centroid for the dissimilar words list and that for the similar words list (e.g., 1.33 – 1.04 = 0.29 for the example lists), called *strength of manipulation on similarity* (SMS), is particularly critical in the context of semantic similarity research given that performance on similar words lists has been contrasted with that on dissimilar words lists in most studies. SMS represents a relative value (dissimilar vs. similar) of the semantic similarity manipulation for a given experiment.

Application of our proposed index to previous studies We have proposed a three-dimensional space and semantic similarity indices, namely, mean distance from the centroid for a list and SMS for an experiment. In fact, applying these indices provides a possible explanation for previous studies. Two classic works with the serial reconstruction task by Baddeley (1966a, 1966b) – the former suggesting a weak effect of semantic similarity and the latter a null effect – have been frequently cited and influential in STM research and thus were selected in particular here. To assess these two studies retrospectively, we calculated the values of our proposed semantic similarity indices for each study. Figure 2 depicts distributions of the mean distances from the centroid for all possible lists by type of list (i.e., similar vs. dissimilar) for Baddeley (1966a), in which five words were randomly drawn from a set of eight words, and two dotted lines representing the mean distance from the centroid for similar and dissimilar words lists from Baddeley (1966b), in which all words in a set were used as a list. Visually, the difference between the mean distance from the centroid of a similar words list and that of a dissimilar words list is generally larger for Baddeley (1966a) than for Baddeley (1966b), suggesting that the manipulation of Baddeley (1966a) was stronger than that of Baddeley (1966b). As a more formal comparison, SMS values (0.56 and 0.34) also imply that Baddeley (1966a) used a manipulation stronger than that of Baddeley (1966b). Furthermore, the distribution of similar words lists overlaps that of dissimilar words lists, indicating that several similar words lists were more dissimilar than the dissimilar words lists in Baddeley (1966a). Thus, the weak effect shown by Baddeley (1966a) could be attributed to these overlaps arising from the variation of stimuli. Therefore, our proposed indices would provide useful information that aids review of the semantic similarity effect in STM.

Critical review of previous studies

We have pointed out that semantic association is a possible confounding factor for semantic similarity and have described CD as an index for the factor of semantic association. We have also proposed that valence, arousal, and dominance are critical dimensions in semantics and that SMS is an index for manipulation on semantic similarity. With the CD and SMS indices, we can address two questions that have not been answered in the literature.

Heterogeneity among studies Some studies using the serial reconstruction task showed a weak detrimental effect of semantic similarity (e.g., Baddeley, 1966a; Crowder, 1979, Exp. 4), while others failed to show it (e.g., Saint-Aubin & Poirier, 1999a, Exp. 3), even though the sample sizes of these studies were not substantially different from each other (N = 20–24), suggesting differences in effect size across studies. For studies using the serial recall task, there were also differences in effect size (see *Results and discussion* section). Thus, there is heterogeneity between studies. As we have

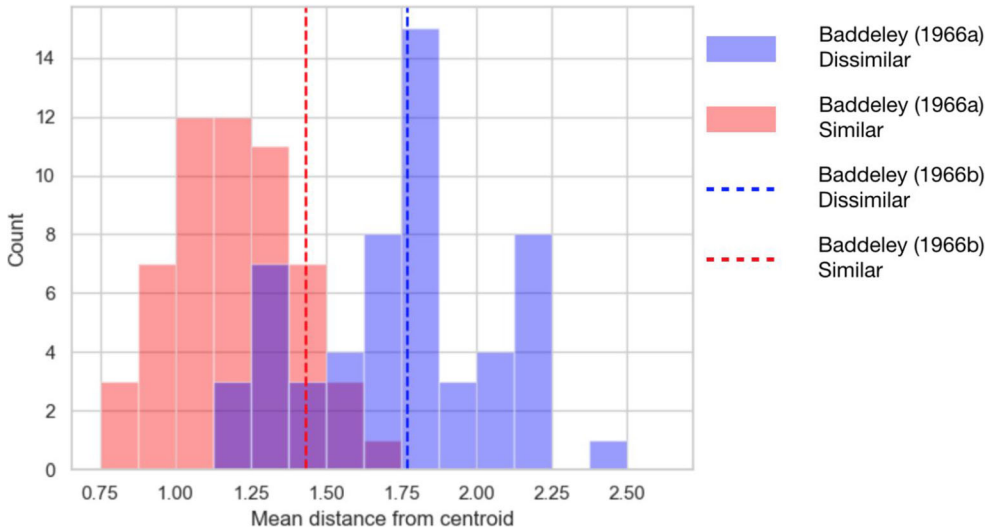

Fig. 2 Values of mean distance from centroid for lists in two previous studies. The histograms represent mean distance from centroid values for similar words lists (colored in red) and dissimilar words lists (colored in blue) in Baddeley (1966a). The red dashed line shows the mean distance from centroid value for the similar words list and the blue dashed line that for the dissimilar words list in Baddeley (1966b). The histogram of similar words lists overlaps that for dissimilar words lists at around 1.5 of the mean distance from centroid value (see also the main text)

proposed SMS as an index for manipulation on similarity, heterogeneity might be attributable to differences in strength of manipulation arising from the selection of materials across studies (i.e., the degree of similarity in similar words lists and dissimilarity in dissimilar words lists).

Facilitative effect of semantic similarity Although STM models generally assume a detrimental effect of similarity (Hurlstone et al., 2014), studies using the serial recall task have shown a facilitative effect of semantic similarity by correct-in-position or item correct scoring (e.g., Poirier & Saint-Aubin, 1995; Saint-Aubin et al., 2005). We infer that semantic association works as a confounding factor, leading to an apparent facilitative effect of semantic similarity. Adopting connectivity difference as an index for the factor of association strength, we assume that a study with a large CD has a large apparent facilitative effect of semantic similarity.

The data from studies using serial reconstruction tasks were reviewed without meta-regression analysis because they are few. For the data from studies using the serial recall task by correct-in-position scoring, we conducted meta-regression analysis with SMS and CD as moderators. Additionally, serial recall data by item correct and conditionalized order errors were reviewed.

Method

Criteria for inclusion in the review

We adopted seven criteria for selecting previous studies. Selected studies (a) reported the materials used in the experiment(s), (b) used single words as stimuli (e.g., not triads of words as in Murdock, 1976), (c) used the serial reconstruction task or the serial recall task, (d) used procedures reflecting one of the three definitions of semantic similarity suggested by Tse (2010), (e) contrasted similar versus dissimilar words list performance, (f) targeted only participants who were adults, and (g) targeted participants who had no cognitive impairment (we included data from the control groups of studies on a patient with brain damage). Most of the studies were selected referring to Tse's (2010) list of previous studies, and other studies that we were aware of were also included. Additionally, to minimize arbitrary selection and the omission of relevant studies, we searched records of all years (1900–2020) included in Web of Science.[7] We also searched for relevant dissertations and theses in ProQuest.[8] The abstracts of all searched studies were checked, and studies that met the criteria were included.

[7] We searched on 21 January 2020 with the following keyword sets (89 hits for keyword set 1 of "semantic similarity, short-term memory, and serial recall"; seven hits for keyword set 2 of "semantic similarity, short-term memory, and reconstruction"; ten hits for keyword set 3 of "semantic relatedness, short-term memory, and serial recall"; and one hit for keyword set 4 "semantic relatedness, short-term memory, and reconstruction").

[8] We searched on 6 May 2020 with the same keyword sets as were used for Web of Science (176 hits for the set 1; 261 hits for the set 2; 74 hits for the set 3; 141 hits for the set 4). Publicly accessible dissertations and theses were targeted due to availability of data.

Creation of possible lists for calculation of indices for previous studies

As values for the two indices – SMS and CD – are not solely determined by individual values of words but are also affected by the composition of lists, possible lists for each study were created for the calculation of the index values for each of the previous studies.[9]

In some previous studies, similar and dissimilar words lists were created by drawing words from two different sets. For instance, Baddeley (1966a) randomly selected five similar words from a similar set of eight words to create a similar words list. Accordingly, we created theoretically possible 56 ($_8C_5$) similar words lists for this study (and likewise, 56 dissimilar words lists were constructed using the dissimilar set). In general, we created possible $_NC_M$ lists, where N is the size of a set and M is the number of words in a list.

In other studies, sets of categorically or associatively grouped words have been used for similar words lists, while dissimilar words lists have been created by drawing words from several different grouped sets. Words on the similar words list were fixed (e.g., words on a list of "animals" did not appear in a list of "fruit"). Thus, grouped sets or $_NC_M$ lists for each grouped set were used as similar words lists. Words on the dissimilar words lists, in contrast, were randomly selected from different grouped sets, which would incur a computational cost when creating possible lists; for example, all combinations of randomized lists of six words from 24 grouped sets of six words would be $_{24}C_6\ 6^6$ = 6,279,710,976. Thus, for computational cost reasons, we randomly selected words from different lists and created 10,000 dissimilar words lists for use in the analysis.

Treatment of non-English materials

Even if non-English words (in our targeted studies, French words) were used in the original experiments, translated English words were used for list construction and subsequent analyses, for four reasons. First, to maintain consistency, the same English norms for valence, arousal, and dominance (Warriner et al., 2013) were referred to for calculating similarity. Second, available French norms for valence and arousal values (Monnier & Syssau, 2014) were based on a relatively small sample (N = 1,031 words) compared with English norms (Warriner et al., 2013) (N = 13,915 words), and consequently, calculating similarity based on French norms was impossible for several French words. We verified the coverage of affective norms using a previous study by Saint-Aubin and Poirier (1999a, Exp. 1), which reported original French words and their counterpart English words. French norms (Monnier & Syssau, 2014) cover only 31% of these French words while English norms (Warriner et al., 2013) cover 83% of these English words. Third, French norms lack dominance values; as we supposed that three dimensions, valence, arousal, and dominance, are important for semantics, we used English norms, for which all three variables are available. Last, valence and arousal values for French words are thought to be acceptably correlated with those for English words. For 42 French–English word pairs reported by Saint-Aubin and Poirier (1999a, Exp. 1), the words were available in norms for the corresponding language. The correlation between valence values of French and English words was r = .78 (95% CI = [0.63, 0.88]) and that between arousal values was r = .73 (95% CI = [0.55, 0.85]). Correlation coefficients based on two randomly formed subgroups of raters of French norms (i.e., inter-rater reliability) were r = .93 for valence and r = .78 for arousal (Monnier & Syssau, 2014); given this, we concluded that inter-language correlations between French and English norms were acceptably high and English norms could substitute for French norms. As we adopted counterpart English words for experiments with French words, we used English free association norms (De Deyne, Navarro, Perfors, Brysbaert, & Storms, 2019) to calculate the CD for each study.

Variables for the review

Strength of manipulation on similarity (SMS) As described in the *Introduction*, we assumed that the semantic information of a word could be expressed as a point in a three-dimensional space defined by valence, arousal, and dominance. The SMS index refers to the difference between the mean Euclidean distance from the centroid of dissimilar words lists and that of similar words list, and it is calculated as follows:

$$\boldsymbol{w}_i = (v_i, a_i, d_i), \tag{1}$$

where $\boldsymbol{w}_i$ represents a point of the ith word of a given list in the semantic space and v, a, and d represent valence, arousal, and dominance values respectively. These values referred to the norms of Warriner et al. (2013). The centroid for a list is calculated as

$$\boldsymbol{c} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{w}_i, \tag{2}$$

where $\boldsymbol{c}$ is the centroid and n is the number of words in a list that appear in the norms (i.e., missing words are omitted). Then, $\boldsymbol{c}$ is calculated when $n \geq 2$; otherwise, $\boldsymbol{c}$ would be identical to $\boldsymbol{w}$ or would not be calculable. The mean distance from the centroid for a list (MD_{list}) is given by the following equation:

[9] We excluded people's names (proper nouns) when constructing the lists because their understood meanings are likely to vary substantially from participant to participant.

$$MD_{list} = \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=1}^{3}\left(w_{ij}-c_j\right)^2\right\}^{\frac{1}{2}}, \quad (3)$$

which represents how dispersedly words are placed in the semantic space (we excluded lists for which the centroid was not calculated). In this equation, j refers to type of dimension (see Eq. 1). As multiple lists are used for an experiment, the values of the mean distance from the centroid are averaged for similar and dissimilar words lists. SMS is given by the following:

$$SMS = \overline{MD}_{dissimilar\ list} - \overline{MD}_{similar\ list}, \quad (4)$$

where $\overline{MD}$ is the averaged value of the mean distance from the centroid for multiple lists. The unique count of words in the materials used from our selected studies was 790, and the English norms for valence, arousal, and dominance (Warriner et al., 2013) covered 87% of these words.

Strength of manipulation on similarity with word2vec (SMS_{w2v}) To contrast our proposed SMS with an index based on a semantic space by a computational model, we also proposed the strength of manipulation on similarity with word2vec (SMS_{w2v}) index based on word2vec (Mikolov et al., 2013). Specifically, we used 300-dimensional vector representations pretrained with Google News corpus (Google, 2013) using a python library, Gensim (Řehůřek & Sojka, 2010) (version 3.7.3). SMS_{w2v} is calculated as follows. As similarity between two words is defined by the cosine of the angle between the two words' vectors in word2vec, word2vec's similarity for a list is given by the following equation:

$$word2vec\ similarity_{list} = \frac{2}{(n^2-n)}\sum_{i=1}^{n}\sum_{j=1}^{n}\cos\left(\boldsymbol{v}_i,\boldsymbol{v}_j\right)\ (i>j), \quad (5)$$

where $\cos(\boldsymbol{v}_i, \boldsymbol{v}_j)$ is the cosine of the angle between two 300-dimensional vectors representing two words of a list ($\boldsymbol{v}_i$ and $\boldsymbol{v}_j$) and n is the number of words on a list that are available in word2vec's 300-dimensional vector representations. When $i = j$, $\cos(\boldsymbol{v}_i, \boldsymbol{v}_j)$ is the angle of the (identical) vectors of the same word. The angle between two vectors is critical for calculation and $\cos(\boldsymbol{v}_i, \boldsymbol{v}_j)$ is interchangeable with $\cos(\boldsymbol{v}_j, \boldsymbol{v}_i)$ because $\cos(\boldsymbol{v}_i, \boldsymbol{v}_j) = \cos(\boldsymbol{v}_j, \boldsymbol{v}_i)$. Thus, we set a constraint $(i > j)$ for calculation. SMS_{w2v} refers to the difference between the mean of word2vec's similarity for similar words lists and that for dissimilar words lists, and is given by the following:

$$SMS_{w2v} = \overline{word2vec\ simialarity}_{similar\ list} - \overline{word2vec\ similarity}_{dissimilar\ list} \quad (6)$$

A model by word2vec with Google News corpus (Google, 2013) covered 98% of the targeted 790 words.

Connectivity difference (CD) English free association norms by De Deyne et al. (2019) provide the response words that participants freely gave to over 12,000 cue words. Based on these norms, a cue-response matrix was created for each list, in which each row represents cue words, each column represents response words, and each cell represents the probability of a cue-response pairing. An example of a cue-response matrix for a list of "apple, banana, orange" is shown in Table 1. Connectivity for a list is defined as the mean cell value except for diagonal cells (see also, Roediger et al., 2001) and is formally defined as the following:

$$connectivity_{list} = \frac{1}{(n^2-n)}\sum_{i=1}^{n}\sum_{j=1}^{n}s_{ij}\ (i\neq j), \quad (7)$$

where n is the number of words in a list that are available in the association norms as cue words and s_{ij} is the associative strength value of the ith row and jth column. For the example list, connectivity is 0.0379 (1/6{0.0203 + 0.0845 + 0.0340 + 0.0136 + 0.0612 + 0.0136}).

Connectivity difference (CD) refers to the difference between the mean connectivity for similar words lists and that for dissimilar words lists, and is given by the following:

$$CD = \overline{connectivity}_{similar\ list} - \overline{connectivity}_{dissimilar\ list}. \quad (8)$$

Most words appear as only response words in the free association norms (De Deyne et al., 2019). For these words, only backward associative strength values (i.e., probability of producing these words as responses when another word is given as a cue) are available but forward associative strength values (i.e., probability of producing words when these words are provided as cues) are missing. By contrast, one word ("anisette") in the association norms appears as only a cue word but not as a response word, which means that participants did not answer "anisette" as a response to over 12,000 cues and the backward associative strength value of that word is calculable as 0: both forward and backward associative strength values are available for this word. Therefore, we targeted words available in the association norms as cue words, for which forward and backward associative strength values are available. Out of all 790 targeted words, 692 words (88%) were available in the association norms as cue words (De Deyne et al., 2019).

Table 1 Example cue-response matrix

Cue	Response		
	Apple	Banana	Orange
Apple		0.0203	0.0845
Banana	0.0340		0.0136
Orange	0.0612	0.0136	

Effect size Studies with a within-subjects design were targeted for effect size calculation because most studies on semantic similarity had within-subjects designs, which can lessen errors due to individual differences. The effect size was defined as per *standardized mean change using change score standardization* (SMCC) (Viechtbauer, 2019), and the mean change score divided by the standard deviation of change scores was estimated. Scores for similar words lists minus those for dissimilar words lists were regarded as change scores. Note that this effect size would be positive when similar words lists' scores are higher than dissimilar words lists' scores and negative when they are lower. Thus, the sign of the effect size represents similarity advantage/disadvantage, while its absolute value represents the standardized size of the effect.

Other variables We also reported design type (within vs. between participants designs), sample size, direction (similarity advantage vs. disadvantage or increase vs. decrease of errors by similarity), reported statistics, and set type (open vs. closed set manipulations). In a closed set manipulation, a limited number of words were repeatedly used across trials, which makes participants familiar with to-be-remembered items (e.g., Baddeley, 1966a). In an open set manipulation, such a constraint is not adopted (e.g., Poirier & Saint-Aubin, 1995).

Procedure

Studies that met the above-mentioned seven criteria were targeted, and three indices (SMS, SMS_{w2v}, and CD) were calculated for each study. In addition to the interpretation of the results of the serial reconstruction task and those of the serial recall task by the three scoring methods in terms of these indices, multiple meta-regression analysis (Harrer, Cuijpers, Furukawa, & Ebert, 2019) was used to analyze the data from studies using the serial recall task by correct-in-position scoring because there were more than nine studies reporting specific statistics. Note that nine is a conventional criterion for the number of studies for meta-regression analysis (see also Harrer et al., 2019). Meta-regression analysis is a regression-based analysis that enables examination of whether moderators (i.e., SMS, SMS_{w2v}, and CD) influence the effect size of each study (i.e., whether there is a semantic similarity effect). Multiple meta-regression was conducted with the metafor package (version 2.1-0) (Viechtbauer, 2019) in R (version 3.5.3) (R Core Team, 2019). A model with SMS (or SMS_{w2v}) as a moderator and a model with both SMS (or SMS_{w2v}) and CD as moderators were examined and then compared by the likelihood ratio test.

Results and discussion

Serial reconstruction task

As only two of nine experiments using the serial reconstruction task reported statistical values, a meta-analysis was not conducted for these studies. Table 2 shows a summary of the results of previous experiments, with the calculated SMS, SMS_{w2v}, and CD variables.

Review by SMS Seven out of the nine studies showed positive SMS values, which suggests that similar words lists of most of the studies were more similar than the dissimilar words lists according to our proposed three-dimensional model. Closer scrutiny revealed there were variations in both SMS values and reported statistics across studies. Two experiments (Baddeley, 1966a; Crowder, 1979, Exp. 4) reported statistically significant detrimental effects of semantic similarity. The former study reported $p < .05$ with the Wilcoxon test and the latter study reported $p < .05$ with a one-tailed ANOVA. These experiments had relatively high SMS values (0.56 and 0.46). In contrast, a study by Nelson et al. (1977) had a negative SMS value (-0.02), which suggests that their similar words lists were *more dissimilar* than their dissimilar words lists in terms of the valence, arousal, and dominance dimensions. Importantly, the results of Nelson and colleagues showed a trend of higher performance for similar words lists than that for dissimilar words lists. Experiments with moderate SMS values (e.g., 0.34 or 0.35) showed both negative and positive trends. Although positive/negative SMS values were generally related to negative/positive directions of an effect, respectively, the results of a neuropsychological case study (Chassé & Belleville, 2009, Exp. 2) did not exhibit this relationship: its SMS was strongly negative (i.e., -0.38) and its trend was also negative. However, as this case study primarily focused on memory performance of a patient with brain damage, the sample size of its control group, whose memory performance was analyzed here, was small ($N = 11$) compared to that of other studies ($N = 20$ or larger). Except for a study with a small sample (Chassé & Belleville, 2009, Exp. 2), the SMS index generally explained differences in the results across studies.

Review by SMS_{w2v} To assess the strength of a semantic similarity manipulation based on a conventional computational model, we also review results along with SMS_{w2v} values. All SMS_{w2v} values were positive, suggesting that the SMS_{w2v} index is indicative of a distinction between similar versus dissimilar words lists. However, it failed to explain differences in the semantic similarity effect across studies. Two studies by Baddeley (1966a, 1966b) had relatively small SMS_{w2v} values (0.13 and 0.10) compared with the other studies, but still showed a significant effect or trend for the detrimental effect. Thus, it is expected that studies with SMS_{w2v}

Table 2 Summary of previous studies using the serial reconstruction task

Study	SMS	SMS_{w2v}	CD	Direction	Statistics	Design	N	Set
Baddeley (1966a, 1)	0.56	0.13	0.024	-1	$p < .05$	Within	21	Closed
Baddeley (1966b, 1)	0.34	0.10	0.025	-1	*n.s.*	Between	40	Open
Chasse & Belleville (2009, 2)	-0.38	0.30	0.010	-1	not reported	Within	11	Closed
Crowder (1979, 3)	0.46	0.24	0.013	-1	*n.s.*	Within	40	Open
Crowder (1979, 4)	0.46	0.24	0.013	-1	$F(1, 19) = 3.04$, $p < .05$	Within	20	Open
Nelson et al. (1977, 2&3)	-0.02	0.22	0.008	1	not reported	Between	96	Open
Saint-Aubin & Poirier (1999a, 3)	0.35	0.30	0.009	1	*n.s.*	Within	24	Open
Saint-Aubin & Poirier (1999a, 4, both conditions)	0.35	0.30	0.009	-1	*n.s.*	Within	56	Closed
Saint-Aubin & Poirier (1999a, 4, first condition)	0.35	0.30	0.009	-1	*n.s.*	Between	56	Closed

Note. In the Direction column, -1 indicates similarity disadvantage while 1 indicates similarity advantage. The Statistics column shows reported statistical values. Baddeley (1966a) used a Wilcoxon test. Baddeley (1966b) focused on learning of a sequence; performance of the first trial was targeted here. Crowder (1979, Exp. 4) used a one-tailed ANOVA. The direction of Nelson et al., (1977) is based on the values reported in the main text (p. 491). They also reported a slight (non-significant) disadvantage for semantic similarity when data were pooled. Saint-Aubin and Poirier (1999a, Exp. 3) used quiet and suppression conditions (a suppression factor); for the semantic similarity effect, these two conditions were collapsed because the interaction between suppression and similarity was not significant. At the test phase of the task in their Experiment 4, words were not presented similarly to the serial recall task; participants, however, had learned all target words thoroughly prior to the experimental session to ensure perfect item recall. Saint-Aubin and Poirier (1999a) designed their Experiment 4 to measure order memory, same as in their Experiment 3 in which they used the serial reconstruction task. They, in fact, noted "[a]s in Experiment 3, subjects were required to remember only order information" (p. 384). Thus, the task of their Experiment 4 is regarded as a variant of the serial reconstruction task

SMS strength of manipulation on similarity, SMS_{w2v} strength of manipulation on similarity based on word2vec, *CD* connectivity difference

values larger than those of Baddeley's studies should show a significant detrimental effect or at least its trend. However, an experiment by Saint-Aubin and Poirier (1999a, Exp. 3) with a SMS_{w2v} value (0.30) being larger than these of Baddeley's two studies (and other studies), showed a trend for a facilitative effect.

Review by CD and set type For the CD index, no clear patterns were identified. Studies with relatively low CD values (Nelson et al., 1977; Saint-Aubin & Poirier, 1999a, Exp.3) showed trends toward similarity *advantage,* which might imply that semantic association impairs memory performance in the serial reconstruction. However, such an interpretation, which is based on semantic association's detrimental effect on serial reconstruction, cannot explain why the study with the largest CD value and a moderate SMS value (Baddeley, 1966b) failed to show a significant effect of similarity disadvantage. For set type, all studies with a closed set showed a significant effect or trend for a detrimental effect. This result might be explained in terms of the item versus order memory distinction: repeating the same words enhances item memory regardless of list type (i.e., similar vs. dissimilar words lists), which offsets a possible facilitative effect of semantic similarity for item memory but accentuates a possible detrimental effect of semantic similarity for order memory.

Summary of data with serial reconstruction task In general, our view that the targeted three dimensions are major determinants of semantic similarity and that semantic similarity has a detrimental effect can explain the results of the previous studies for the SMS index. According to the SMS values, when the strength of manipulation on similarity is positive and large, there is likely to be a detrimental effect of semantic similarity. Furthermore, the negative value of SMS for Nelson et al. (1977) can explain this study's trend of advantage for similar words lists: even if lists were assumed to be composed of similar words, they could be more dissimilar than the counterpart lists of dissimilar words. The SMS index explains the current data more appropriately than the SMS_{w2v} index does. Therefore, it was concluded that the SMS index is useful for interpreting results from previous studies.

Serial recall task by correct-in-position scoring

The results of experiments with the serial recall task are summarized in Table 3. Most results (18 out of 22) showed statistical significance for or trends toward a semantic similarity advantage, consistent with the common view that semantic similarity works facilitatively for serial recall by correct-in-position scoring.[10] The results of 14 studies, for which the effect size was calculable, were targeted for meta-analysis. Figure 3 is a forest plot representing the estimated effect size (i.e., standardized mean difference) and total effect size.

[10] It is problematic to review previous studies with set type because studies using closed sets had small sample sizes ($N = 4 - 11$) and did not report specific statistics. Thus, we did not address set type differences for data by correct-in-position scoring. This is true for data scored by item correct or conditionalized order errors.

Table 3 Summary of previous studies using the serial recall task by correct-in-position scoring

Study	SMS	SMS_{w2v}	CD	Direction	Statistics	Design	N	Set
Belleville et al. (2003, 4)	-0.48	0.24	0.001	1	not reported	Within	4	Closed
Biegler (2007, 7, Closed)	0.37	0.26	0.010	-1	not reported	Within	8	Closed
Biegler (2007, 7, Open)	0.37	0.26	0.010	1	not reported	Within	8	Open
Chasse & Belleville (2009, 1)	0.05	0.18	0.002	1	not reported	Within	10	Closed
Chasse & Belleville (2009, 2)	-0.38	0.30	0.010	1	not reported	Within	11	Closed
Crowder (1979, 1)	0.27	0.30	0.003	unclear	*n.s.*	Between	24	Open
Crowder (1979, 2)	0.65	0.19	0.023	unclear	$F(1, 30) = 1.04$, $p = .316$	Between	32	Open
Crowder (1979, 5)	0.45	0.24	0.013	1	$t(19) = -.15$, $p = .882$	Within	20	Open
Guérard & Saint-Aubin (2012, 3)	0.33	0.30	0.009	1	$F(1, 19) = 59.60$, $p < .001$	Within	20	Open
Hadley (2006, 4)	0.33	0.33	0.008	1	$F(1, 46) = 32.629$, $p < .001$	Within	48	Open
Nelson et al. (1977, 4)	-0.02	0.22	0.008	-1	*n.s.*	Between	64	Open
Poirier & Saint-Aubin (1995, 1)	0.11	0.29	0.005	1	$F(1, 23) = 49.37$, $p < .001$	Within	24	Open
Poirier & Saint-Aubin (1995, 2)	0.11	0.29	0.005	1	$F(1, 15) = 37.28$, $p < .001$	Within	16	Open
Poirier & Saint-Aubin (1995, 3)	0.16	0.31	0.008	1	$F(1, 15) = 20.14$, $p < .001$	Within	16	Open
Saint-Aubin & Poirier (1999a, 1)	0.35	0.31	0.011	1	$F(1, 23) = 39.70$, $p < .001$	Within	24	Open
Saint-Aubin & Poirier (1999a, 2)	0.35	0.30	0.009	1	$F(1, 23) = 3.75$, $p = .065$	Within	24	Open
Tse (2009, Mixed-associative)	0.44	0.24	0.035	1	$t(19) = 6.41$, $p < .001$	Within	20	Open
Tse (2009, Mixed-categorical)	0.46	0.33	0.013	1	$t(19) = 2.17$, $p < .05$	Within	20	Open
Tse (2009, Pure-associative)	0.44	0.24	0.035	1	$t(24) = 5.64$, $p < .001$	Within	25	Open
Tse (2009, Pure-categorical)	0.46	0.33	0.013	1	$t(24) = 1.91$, $p = .068$	Within	25	Open
Tse et al. (2011, Associative)	0.44	0.24	0.035	1	$t(151) = 9.50$, $p < .001$	Within	152	Open
Tse et al. (2011, Categorical)	0.46	0.33	0.013	1	$t(151) = 6.73$, $p < .001$	Within	152	Open

Note. In the Direction column, -1 means similarity disadvantage while 1 means similarity advantage. Poirier and Saint-Aubin's experiments (Poirier & Saint-Aubin, 1995, Exps. 2 and 3; Saint-Aubin & Poirier, 1999a, Exps. 1 and 2) included a suppression factor (quiet and suppression conditions), but only one of these four experiments showed an interaction between semantic similarity and suppression (Saint-Aubin & Poirier, 1999a, Exp. 2); thus, quiet and suppression conditions were collapsed

SMS strength of manipulation on similarity, SMS_{w2v} strength of manipulation on similarity based on word2vec, *CD* connectivity difference

Although between-study heterogeneity was large ($I^2 = 99\%$), the total effect size (0.90) and prediction intervals (95% PI = [-0.20, 1.99]) generally support a similarity advantage. Egger's test (Egger, Smith, Schneider, & Minder, 1997) did not detect

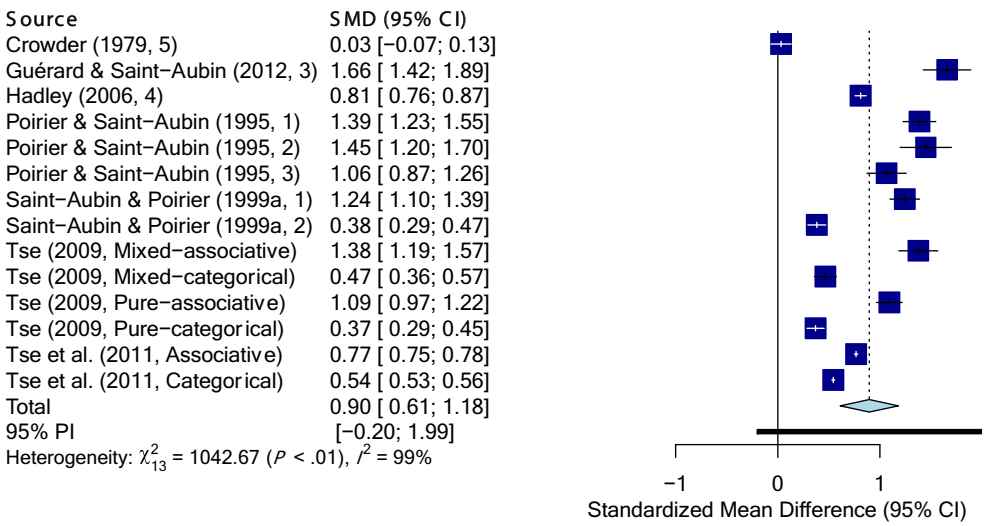

Fig. 3 Forest plot for previous studies on the semantic similarity effect with the serial recall task. SMD refers to standardized mean difference (standardized difference between performance for similar and for dissimilar words lists)

a significant intercept (t = 1.12, p = 0.29), and it was therefore concluded that the overall effect was not substantially biased by studies with a small sample size but a large effect size.

Although the results of the meta-analysis appear to suggest a facilitative semantic similarity effect, the results should be interpreted as showing that manipulations by categorically and/or associatively grouped lists, per se, have facilitative effects. Thus, meta-regression was conducted to examine two possibilities: (a) semantic similarity as such has a detrimental effect and (b) a possible confounding factor, semantic association, has a facilitative effect. Model 1, with the semantic similarity index (either SMS or SMS_{w2v}) as a moderator, and Model 2, with the semantic similarity index (either SMS or SMS_{w2v}) and semantic association index (CD) as moderators, were tested.

Model 1 with semantic similarity index To examine the semantic similarity effect in terms of a semantic similarity index, SMS was first used as a semantic similarity index. Model 1 with SMS explains R^2 = 34.00% of the heterogeneity in our data and indicates that the SMS value has a negative effect on effect size (t = -2.41, p = .0328) (Table 4). This suggests that semantic similarity has a detrimental effect on serial recall, because effect size is a standardized mean difference between performance on similar words lists and that on dissimilar words lists. As the strength of manipulation on similarity increases, the advantage of similar words lists over dissimilar words lists decreases and, theoretically, turns to a disadvantage.

Second, SMS was replaced with SMS_{w2v} and Model 1 was tested. Model 1 with SMS_{w2v} accounts for only a small amount of the heterogeneity (R^2 = 1.93%). The direction implies the detrimental effect of semantic similarity based on SMS_{w2v} but it did not reach statistical significance (t = -0.48, p = .6427) (Table 4).

Relationship between semantic similarity and semantic association indices Note that a positive value of SMS or SMS_{w2v} supports the idea that similar words lists are, as intended, more similar than dissimilar words lists, and a positive value of CD implies that similar words lists are more associated than dissimilar words lists. All SMS, SMS_{w2v}, and CD values for the selected studies in this meta-analysis showed positive values, indicating that similar words lists were more associated than dissimilar words lists. Furthermore, the SMS values for each study were correlated with CD values (r = .57), suggesting that a study with a strong manipulation on semantic similarity also had a strong manipulation on semantic associations. In contrast, SMS_{w2v} values negatively correlated with CD values (r = -.73), which is difficult to interpret but may question the utility of the SMS_{w2v} index. As was mentioned in the interpretation of data with the serial reconstruction task, SMS_{w2v} may be useful for distinguishing similar words lists from dissimilar words lists but not for explaining differences across studies.

The relationship between CD and SMS or SMS_{w2v} suggests that semantic association might have been a confounding factor within single studies. In addition, the correlation between CD and SMS implies that semantic association might

Table 4 Results of four models in the meta-regression analysis

	Estimate	SE	t	p	R^2
Model 1 with SMS					34.00%
Intercept	1.66	0.338	4.91	.0004***	
SMS	-2.19	0.907	-2.41	.0328*	
Model 1 with SMS_{w2v}					1.93%
Intercept	1.41	1.088	1.29	.2199	
SMS_{w2v}	-1.75	3.687	-0.48	.6427	
Model 2 with SMS and CD					53.33%
Intercept	1.71	0.305	5.62	.0002***	
SMS	-3.31	0.986	-3.36	.0064**	
CD	22.45	11.227	2.00	.0708†	
Model 2 with SMS_{w2v} and CD					3.19%
Intercept	2.01	1.874	1.07	.3075	
SMS_{w2v}	-3.40	5.622	-0.60	.5576	
CD	-7.62	19.107	-0.40	.6975	

Note. SMS is calculated based on our proposed valnece-arousal-dominance space while SMS_{w2v} based on a semantic spance provided by word2vec
†p < .10. *p < .05. **p < .01. ***p < .001

SMS strength of manipulation on similarity, *SMS_{w2v}* strength of manipulation on similarity based on word2vec, *CD* connectivity difference

have been confounded with semantic similarity across studies. Thus, to control for this association, we added both the semantic similarity index (SMS or SMS_{w2v}) and the semantic association index (CD) variables to Model 2.

Model 2 with both semantic similarity and semantic association indices Model 2 with SMS and CD shows that R^2 accounts for 53.33% of the heterogeneity and again suggests a detrimental SMS effect (t = -3.36, p = .0064) and a trend toward a facilitative CD effect (t = 2.00, p = .0708) (Table 4). Although the CD effect failed to reach statistical significance, a likelihood ratio test favors Model 2 over Model 1 (χ^2 = 4.3960, p = .0360), suggesting that both SMS and CD serve as moderators of the semantic similarity effect. Note that the coefficient of the SMS term is more negative for Model 2 than for Model 1, supporting the assertion that controlling for semantic association provides clear evidence for the detrimental effect of semantic similarity.

For comparison purposes, Model 2 with SMS_{w2v} and CD was also checked. This model explains little of the heterogeneity across studies (R^2 = 3.19%). It suggests a direction toward a detrimental SMS_{w2v} effect (t = -0.60, p = .5576) and a detrimental CD effect (t = -0.40, p = .6975), but they are not statistically significant (Table 4). Because of the lack of statistical evidence, this seemingly unexpected result of a direction for a detrimental effect of CD (i.e., semantic association) is difficult to interpret. We did not compare Model 1 with Model 2 using the SMS_{w2v} index because of the small percentage of explained heterogeneity.

Intercept of Model 1 and Model 2 Models with the SMS index seem to provide a clear explanation for previous studies. Nevertheless, the intercepts of both Model 1 with the SMS index and Model 2 with the SMS and CD indices were positive values (1.66 and 1.71). Thus, even if semantic similarity and connectivity were equated for similar and dissimilar words lists (i.e., a hypothetical case in which SMS = 0 and CD = 0), a facilitative effect would remain. This implies that manipulations of previous studies had a facilitative effect for similar words lists that cannot be explained by semantic association (or semantic similarity). We infer that the effect that cannot be attributed to semantic association (or semantic similarity) would be the effect of the extracted category or theme label's cuing (see the *General discussion* section).

Summary of data with serial recall task by correct-in-position A model with SMS and CD indices explained a considerable percentage of heterogeneity across studies (R^2 = 53.33 %), which suggests heterogeneity across studies is partly attributable to differences in semantic similarity and semantic association across studies. More importantly, results of this model indicated a detrimental effect of semantic similarity and a facilitative effect of semantic association for serial recall by correct-in-position scoring. Although SMS_{w2v} was indicative of the distinction between similar versus dissimilar words lists, it failed to explain the differences across studies.

Serial recall task by other scorings

In the following sections, we review data by item correct and conditionalized order errors scoring. The number of studies with a within-participants design that reported statistical values was four for item correct data and six for conditionalized order errors data. Given the small number of studies, we did not conduct a meta-regression.

Item correct scoring All but one study converged to show a significant effect for or a trend toward the facilitative effect of semantic similarity on serial recall performance by item correct scoring (Table 5). The exception (Nelson et al., 1977) is the only study with a negative SMS value (and the second smallest SMS_{w2v} value) among the 14 studies, which might imply that semantic similarity is facilitative to item memory. However, other potential explanations are conceivable. For example, semantic association might have contributed to the apparent semantic similarity advantage because all studies reporting a semantic similarity advantage had positive CD values. Alternatively, a label's cuing might have affected the apparent semantic similarity advantage.

Conditionalized order errors scoring Table 6 shows the results by conditionalized order errors scoring. Five out of eight studies showed a significant effect or a trend implying that semantic similarity increases conditionalized order errors and these five studies had relatively large SMS values (≥0.34). This result might be explained by assuming a detrimental effect of semantic similarity on order memory. Although the SMS index cannot explain why a study with a large negative SMS value (Chassé & Belleville, 2009, Exp. 2) did not show the decrement of order errors, it might be because of its limited sample size (N = 11). For the SMS_{w2v} index, no clear patterns were identified. Studies with relatively large CD values (≥0.011) showed increases in conditionalized order errors, which would indicate a detrimental effect of semantic association on order memory (Poirier et al., 2015), but they also had relatively large SMS values (0.34 or 0.45). Although such a correlation between CD and SMS indices makes interpreting patterns of results difficult in terms of separating semantic association and semantic similarity, comparing two studies with the same CD value (CD = 0.009) (Guérard & Saint-Aubin, 2012, Exp. 3; Saint-Aubin & Poirier, 1999a, Exp. 2) suggests that semantic similarity can explain differences in the results well given that the size of the SMS value of these two studies (0.33 and 0.35) corresponds to a decrease/increase of conditionalized order errors (-1 and 1). The overall pattern of results supports the detrimental effect of semantic similarity

Table 5 Summary of previous studies using the serial recall task by item correct scoring

Study	SMS	SMS_{w2v}	CD	Direction	Statistics	Design	N	Set
Biegler (2007, 7, Closed)	0.38	0.26	0.010	1	not reported	Within	8	Closed
Biegler (2007, 7, Open)	0.38	0.26	0.010	1	not reported	Within	8	Open
Crowder (1979, 1)	0.27	0.30	0.003	1	$F(1,22) = 4.67, p < .05$	Between	24	Open
Crowder (1979, 2)	0.65	0.19	0.023	1	$F(1,30) = 19.65, p < .001$	Between	32	Open
Crowder (1979, 5)	0.45	0.24	0.013	1	not reported	Within	20	Open
Hadley (2006, 4)	0.34	0.33	0.008	1	$F(1, 46) = 122.567, p < .001$	Within	48	Open
Nelson et al. (1977, 4)	-0.02	0.22	0.008	-1	*n.s.*	Between	64	Open
Poirier & Saint-Aubin (1995, 1)	0.11	0.29	0.005	1	$F(1, 23) = 82.92, p < .001$	Within	24	Open
Poirier & Saint-Aubin (1995, 2)	0.11	0.29	0.005	1	$F(1, 15) = 152.92, p < .001$	Within	16	Open
Poirier & Saint-Aubin (1995, 3)	0.16	0.31	0.008	1	$F(1, 15) = 52.87, p < .001$	Within	16	Open
Tse (2009, Mixed-associative)	0.44	0.24	0.035	1	not reported	Within	20	Open
Tse (2009, Mixed-categorical)	0.46	0.33	0.013	1	not reported	Within	20	Open
Tse (2009, Pure-associative)	0.44	0.24	0.035	1	not reported	Within	25	Open
Tse (2009, Pure-categorical)	0.46	0.33	0.013	1	not reported	Within	25	Open

Note. In the Direction column, -1 means similarity disadvantage while 1 means similarity advantage. Tse (2009) contrasted the amounts of increase for item correct by associatively or categorically grouping using item correct for ungrouped lists as a baseline and showed associatively grouping led to an increase larger than categorically grouping. Specific statistics for associatively or categorically grouping vs. baseline are, however, not reported

SMS strength of manipulation on similarity, SMS_{w2v} strength of manipulation on similarity based on word2vec, *CD* connectivity difference

on order memory but does not show firm evidence for the detrimental effect of semantic association on order memory separately from the assumed semantic similarity effect.

General discussion

STM models generally suppose a detrimental effect of the similarity of stimulus properties (Hurlstone et al., 2014), which is supported by demonstrations of phonological (Baddeley et al., 1984), tonal (Williamson et al., 2010), and visual (Avons & Mason, 1999) similarity effects. By contrast, observations of the semantic similarity effect suggest weak or null detrimental effects for serial reconstruction and even facilitative effects for serial recall by correct-in-position scoring, implying an inconsistency between the semantic similarity effect and the effect expected by STM models.

The facilitative effect for serial recall has been addressed by using item correct and conditionalized order errors scorings that are assumed to measure item and order memory, respectively. With both scoring methods, previous studies have implied that semantic similarity is facilitative to item memory but detrimental or neutral to order memory (e.g., Saint-Aubin et al., 2005). As the serial reconstruction task is thought to measure order memory (Saint-Aubin & Poirier, 1999b), the results of a detrimental effect of semantic similarity in this task might show a detrimental similarity effect on order memory. Nevertheless, as noted, the detrimental effect of semantic

Table 6 Summary of previous studies using the serial recall task by conditionalized order errors scoring

Study	SMS	SMS_{w2v}	CD	Direction	Statistics	Design	N	Set
Chasse & Belleville (2009, 1)	0.05	0.18	0.002	-1	not reported	Within	10	Closed
Chasse & Belleville (2009, 2)	-0.38	0.30	0.010	0	not reported	Within	11	Closed
Guérard & Saint-Aubin (2012, 3)	0.33	0.30	0.009	-1	$F(1,19) = 1.30, p = 0.27$	Within	20	Open
Saint-Aubin & Poirier (1999a, 1)	0.34	0.31	0.011	1	$F(1, 23) = 4.09, p = .06$	Within	24	Open
Saint-Aubin & Poirier (1999a, 2)	0.35	0.30	0.009	1	$F(1, 23) = 2.10, p = .16$	Within	24	Open
Tse (2009, Mixed)	0.45	0.29	0.024	1	$F(1, 19) = 17.06, p < .001$	Within	20	Open
Tse (2009, Pure)	0.45	0.29	0.024	1	$F(1, 48) = 8.17, p < .01$	Within	50	Open
Tse et al. (2011)	0.45	0.29	0.024	1	$F(1, 150) = 11.74, p < .001$	Within	152	Open

Note. In the Direction column, -1 means a decrease of order errors by semantic similarity while 1 means an increase of order errors by semantic similarity

SMS strength of manipulation on similarity, SMS_{w2v} strength of manipulation on similarity based on word2vec, *CD* connectivity difference

similarity for serial reconstruction is weak or even null in contrast to a robust detrimental effect of phonological similarity for serial reconstruction (e.g., Baddeley, 1966a). Furthermore, given that the detrimental effect of phonological similarity has been demonstrated for serial recall even by correct-in-position scoring (e.g, Watkins et al., 1974), it is valuable to examine the cause of this unique influence of semantic similarity manipulations.

In the present study, we took an alternative approach to the item versus order memory distinction. Specifically, we aimed to quantify semantic similarity and identify a possible confounding factor. This approach does not challenge the item versus order memory distinction but rather clarifies the semantic similarity effect even within the framework based on the item versus order memory distinction. We reviewed studies on the semantic similarity effect in STM using our proposed indices for semantic similarity and semantic association, and used these indices to conduct a meta-regression analysis. The SMS index represents the manipulation strength of semantic similarity, based on the assumption that semantic similarity is viewed as spatial proximity in the valence-arousal-dominance semantic space. Additionally, for comparison purposes, the SMS_{w2v} index, which refers to the manipulation strength of semantic similarity defined by word2vec, was also used. The CD index represents the inter-item associative strength of similar words lists relative to that of dissimilar words lists. The findings are summarized below.

As the total effect size suggests, manipulations used in previous studies on semantic similarity had a facilitative effect for serial recall by correct-in-position scoring. Although these manipulations, such as categorically grouping words, were developed to manipulate semantic similarity, they also affected factors other than semantic similarity; in fact, it was suggested that the semantic similarity component of these manipulations had a detrimental effect on both serial recall by correct-in-position scoring and serial reconstruction. A possible confounding factor, semantic association, was also examined and shown to have a facilitative effect for serial recall by correct-in-position scoring. In addition, a review of data by item correct and conditionalized order errors scorings would suggest that semantic association is facilitative to item memory while semantic similarity is detrimental to order memory; studies showing an advantage of categorically or associatively grouped lists by item correct scoring had positive CD values while studies showing increases of conditionalized order errors had large SMS values.

Index for strength of manipulation on similarity

The semantic similarity effect on STM has typically been investigated by testing differences between STM performance on similar words lists and that on dissimilar words lists. Thus, experimental results are likely dependent on the selection of materials for the respective lists. More specifically, the semantic similarity strength of selected similar words lists relative to that of selected dissimilar words lists (i.e., manipulation strength) should influence the difference in STM performance between these two types of lists. Therefore, even null results from an experiment using the serial reconstruction task cannot be taken to show the absence of a semantic similarity effect unless the possibility of an insufficient manipulation of semantic similarity is ruled out. We proposed the SMS index as quantification of manipulation strength of semantic similarity and demonstrated that a study with a large manipulation strength tends to show a large detrimental effect of semantic similarity for both serial reconstruction and serial recall by correct-in-position. Given that serial reconstruction is thought to reflect order memory, the detrimental effect of semantic similarity is likely to affect order memory. Furthermore, as studies with relatively large SMS values show increased conditionalized order errors, the detrimental effect of semantic similarity for serial recall would also be attributable to the semantic similarity effect on order memory. We suggest that quantification of semantic similarity (e.g., SMS index) is desirable even for a single study, because manipulations of categorical or associative groupings would not necessarily assure semantic similarity. For example, a study by Nelson et al. (1977) did not show a significant effect in which similar words lists were more dissimilar than dissimilar words lists according to the SMS index.

As we regard spatial proximity in semantic space as tantamount to similarity, we argue that semantic similarity of a list is not binary but rather a continuum; consequently, a similar words list can be very similar, moderately similar, or barely similar. This further implies that a comparison of different types of similarity should use equivalently similar materials, that is, the comparison should quantitatively match the degree of similarity, echoing Huttenlocher and Newcombe's (1976) idea that "a valid comparison of the effects of acoustic and semantic similarity would require some metric for equating the degree of acoustic and semantic similarity" (p. 392).

Index for connectivity difference

In the literature, "semantic association" is often used interchangeably with "semantic similarity." Moreover, semantic association is known to have a facilitative effect on STM performance (e.g., Saint-Aubin et al., 2014), and it might act as a confounding factor for semantic similarity. In previous studies' settings, semantic association seems to be correlated with semantic similarity: a study with a large SMS value tends to have a relatively large CD. Given the relation of semantic similarity to semantic association in previous studies and given semantic association's influence on STM performance, researchers should consider whether semantic association affects STM performance in studies on the semantic similarity

effect. The results of a meta-regression suggest that semantic association indeed has a positive effect; therefore, we suggest that semantic association contributes to an apparent facilitative effect of semantic similarity. As patterns of CD do not correspond to patterns of effects for serial reconstruction, semantic association would affect item memory but not order memory.

Facilitative effects of manipulations in previous studies

Nevertheless, association alone cannot explain fully the facilitative effects of manipulations in previous studies, because the intercepts of the models by meta-regression have positive values. The presence of a facilitative effect that is not attributable to association is consistent with the notion that manipulations in previous studies, such as categorical or associative groupings, have effects above and beyond those of semantic similarity or semantic association. For categorically or associatively grouped lists (i.e., similar words lists under the associative or categorical relatedness definitions), participants would internally generate category and/or theme labels and use these labels as retrieval cues (Poirier & Saint-Aubin, 1995; Saint-Aubin et al., 2005; Saint-Aubin & Poirier, 1999a; Tse, 2009).

Our explanation for the semantic similarity effect, with future directions

Based on the above reasoning, we propose a unified explanation of the results of previous studies on the semantic similarity effect in STM. For serial reconstruction, as all items are available at retrieval, the influence of categorical or theme labels as retrieval cues is likely to be limited (Saint-Aubin & Poirier, 1999a). Consequently, a detrimental effect of semantic similarity has been observed in some studies (Baddeley, 1966a; Crowder, 1979). However, demonstrating the semantic similarity effect depends on the manipulation strength for semantic similarity, and a study would fail to demonstrate the effect if its manipulation strength is too small. For serial recall, items are not presented at retrieval, and so the facilitative effect of retrieval cues for retrieving items would be substantial. As associative or categorical grouping allows such retrieval cues to affect memory performance (e.g., Saint-Aubin et al., 2005; Tse, 2009), the previous studies show the advantage of associatively or categorically grouped lists, which has been assumed to be due to the facilitative effect of semantic similarity. Nevertheless, as the results of meta-regression suggest, semantic similarity has a detrimental effect on memory, and the facilitative effect of retrieval cues leads to an apparent facilitative effect of semantic similarity. Given observations of the detrimental effect of semantic similarity for the serial reconstruction task and increases of conditionalized order errors by semantic similarity in serial recall, we assume that semantic similarity leads to confusion between items, which can be observed as order errors. In addition, semantic association, which seems confounded with similarity in the settings of previous studies, also contributes to the apparent facilitative effect for retrieving items; our results imply that semantic association has a facilitative effect for serial recall by correct-in-position and possibly for item correct scorings but it does not affect serial reconstruction. In light of the item versus order memory distinction, our explanation supposes that semantic similarity is neutral to item memory but detrimental to order memory while semantic association is facilitative to item memory but neutral to order memory. Furthermore, the effect of retrieval cues for item memory depends on whether a task requires the retrieval of items (e.g., the serial recall task) or not (e.g., the serial reconstruction task). According to our explanation, the facitiative effect of semantic similarity on item memory that is supposed by a common view (e.g., Saint-Aubin et al., 2005) would be explained by the facilitative effects of semantic association and additional retreival cues on item memory (i.e., confounding factors in previous studies' settings).

The core of our explanation is consistent with existing STM models (Brown et al., 2007; Farrell, 2006; Nairne, 1990; Page & Norris, 1998) and with observations of the detrimental effect of similarity on STM with a variety of stimulus properties, such as phonological (Baddeley et al., 1984), tonal (Williamson et al., 2010), and visual information (Avons & Mason, 1999). Our explanation, however, seems contradictory to the findings by Poirier et al. (2015), who showed that presenting semantically associated words increased order errors. In their first experiment, memory performance for experimental lists and control lists of six words was examined. For an experimental list (e.g., "band, record, concert, yellow, music, tourist"), the first three words (e.g., "band," "record," and "concert") were semantically associated with the target fifth word (e.g., "music"), defined by associative strength of free association norms (Nelson et al., 2004). In contrast, for a control list (e.g., "band, record, concert, tractor, fence, police"), the first three words (e.g., "band," "record," and "concert") were not associated with the target fifth words (e.g., "fence"). The fourth and sixth words were filler words that were unrelated to other words in both the experimental and control lists. The results showed that recalling of the target fifth word regardless of its position (i.e., item correct) was higher for experimental lists than for control lists although recalling of the target fifth word at its correct position (i.e., correct-in-position) was equivalent for experimental and control lists. As the difference between item correct score and correct-in-position score reflected the number of words recalled at wrong positions, their subsequent analysis on order errors demonstrated that the target fifth words were recalled at wrong positions more frequently for experimental lists than for control lists. Theoretically, Poirier et al. (2015) suggest that the

primacy gradient (e.g., Page & Noris, 1998), which encodes the activation level of items' representations and also represents order information, can be interpreted as the activation level of items' representations within a semantic associative network. Therefore, spreading activation in a semantic associative network caused by encoding semantically associated words leads to disturbance in the primacy gradient patterns, which is observed as order errors. The results and explanation of Poirier et al. (2015) can also be viewed as showing a form of interplay between item and order memory instead of a clear-cut distinction between them.[11]

Nevertheless, our explanation based on the assumption that semantic similarity is detrimental to order memory while semantic association is facilitative to item memory would explain the findings by Poirier et al. (2015). First, we note that semantic similarity may act as a confounding factor for semantic association in a study on semantic association and that the increase of order errors for experimental lists in Poirier et al. (2015) could be attributable to a detrimental effect of experimental lists' semantic similarity on order memory. In fact, in the current study, we have suggested that (seemingly) detrimental effects of semantic association on order memory in serial reconstruction and serial recall scored by conditionalized order errors can be explained by a detrimental effect of semantic similarity on order memory. Second, we suggest that a higher item correct score for the target fifth words in the experimental lists of Poirier et al. (2015) would reflect the facilitative effect of semantic association on item memory. Third, as serial recall by correct-in-position score is thought to measure both item and order memory, it is possible that semantic similarity's detrimental effect (for order memory) offsets semantic association's facilitative effect (for item memory) in correct-in-position score, leading to equivalent correct-in-position scores for the experimental and control lists of Poirier et al. (2015). At the least, results of our meta-regression suggest a detrimental effect of semantic similarity and a facilitative effect of semantic association on serial recall scored by correct-in-position. Thus, these effects may counteract with each other. Accordingly, our explanation is not incompatible with the findings by Poirier et al. (2015) and it implies the theoretical importance of the distinction between semantic association and semantic similarity in STM studies on semantics.

Our explanation does not merely fit STM models and observations of similarity effects, but can also offer a novel prediction. Specifically, if the following conditions are met: (a) influence of categorical or theme labels is minimized for similar and dissimilar words lists, (b) semantic association is equated for these two types of lists, and (c) manipulation strength is large enough, then memory performance for similar words lists is predicted to be lower than that for dissimilar words lists in the serial recall task even by correct-in-position scoring. A feasible mode of construction for such a similar words list would be drawing words from different categories or sets of theme-related words, the same as in the construction of dissimilar words lists, while ensuring that the valence, arousal, and dominance values of words are similar to each other and there is equal strength of semantic association (i.e., connectivity) for similar and dissimilar words lists. Given that the influence of categorical or theme labels on memory performance is recognized in this context, further research with this alternative approach is needed as a direct test of the semantic similarity effect, experimentally controlling for the effect of labels and semantic association.

In this study, we targeted the effect size of each study because memory scores by list and study were not available, but it would be desirable to test how semantic similarity and semantic association affect memory performance at the list level by applying linear regression analysis to memory scores for each list, with semantic similarity and semantic association indices for each list as explanatory variables (see also Roediger et al., 2001). Based on our explanation, semantic similarity is expected to have a detrimental effect even at the list level, and semantic association a facilitative effect.

Affect in semantics

Our assumption that the valence, arousal, and dominance dimensions play an important role in semantically imbued behaviors (i.e., STM performance) was directly drawn from Osgood and associates' view of semantics (e.g., Miron, 1969; Osgood & Suci, 1955) as well as recent findings from computational natural language processing research (e.g., Bestgen & Vincze, 2012; Hollis & Westbury, 2016). This assumption is consistent with evidence showing that affective values of words influence cognitive performance, such as lexical decision and naming (e.g., Kousta, Vinson, & Vigliocco, 2009; Kuperman, Estes, Brysbaert, & Warriner, 2014). Research also suggests that release from proactive interference (PI), a memory phenomenon, occurs when words shift from one pole to the other pole of an affective dimension (e.g., a trial of high-valence words to a trial of low-valence words)

[11] Another key finding by Poirier et al. (2015, Exps. 1 & 2) was that when items were recalled at wrong positions, they tended to be recalled early (i.e., *anticipation errors*) more than late (i.e., *postponement errors*) and that anticipation errors were observed more in experimental lists than in control lists while postponement errors were equivalent for the two types of lists. These results were, in fact, predicted by the view that activation levels in a semantic network represent order memory, and thus they were thought to support this view. Nevertheless, it should be noted that the preponderance of anticipation errors over postponement errors was observed even in their control lists. The preponderance of anticipation errors in immediate serial recall was also reported by other studies that did not manipulate a semantic factor (Haberlandt, Thomas, Lawrence, & Krohn, 2005; Ma et al., 2019). Given the baseline tendency toward anticipation errors, any factor that causes order errors may accentuate anticipation errors. We believe that further research should compare effects of a semantic factor and another factor (e.g., phonological similarity) on transposition gradients to examine whether a semantic factor increases anticipation errors in particular.

(Wickens & Clark, 1968) and that the size of the effect of release from PI has a linear relationship with the Euclidean distance made by the shift in the affectively defined semantic space (Weeks, 1976). These findings indicate the influence of affective dimensions on memory and further support our assumptions that the values of affective dimensions are continuous and that semantic similarity can be viewed as spatial proximity in the space of affective dimensions.

More importantly, results of previous studies have shown the effect of emotions on STM but also implied complex patterns for their effect, regarding differences in positive, neutral, and negative emotion words (Majerus & D'Argembeau, 2011; Monnier & Syssau, 2008; Tse & Altarriba, 2009). For example, Monnier and Syssau (2008) demonstrated that positive emotion words were recalled more than neutral words for immediate serial recall. In contrast, Tse and Altarriba (2009) did not show an advantage of positive emotion words over neutral words but they showed a *disadvantage* of negative emotion words over neutral words in the immediate serial recall task. As semantic relatedness (i.e., semantic association) is thought to be one factor leading to emotion words' advantage for memory, with the assumption that emotion words are related to each word (Talmi, Luk, McGarry, & Moscovitch, 2007; Talmi & Moscovitch, 2004), Majerus and D'Argembeau (2011) controlled semantic relatedness for positive, negative, and neutral words lists and conducted the immediate serial recall task. Throughout their four experiments, the advantage of positive words was repeatedly demonstrated but the advantage of negative words was partly supported (e.g., item errors were higher for neutral words than for negative words in Exp. 1, but this difference was not replicated in Exp. 2). Thus, despite the complex patterns of results regarding trifurcation of positive, neutral, and negative emotions, previous studies have provided evidence for the effect of emotion on STM.

It should be noted that Majerus and D'Argembeau (2011) propose a theoretical position that the emotional-semantic space covers a part of the whole lexico-semantic space, which is different from our position that affective dimensions underlie the entirety of semantics based on Osgood and associates' view (e.g., Miron, 1969; Osgood & Suci, 1955). We infer that this difference in views derives from the difference in approaches to assessing emotion in semantics. STM studies on emotion have typically contrasted extremely positive/negative words with neutral words by assuming the trifurcation of positive, neutral, and negative words, whereas our present study targets the distribution of affective values in common words. In the former approach, it would be theoretically natural to compartmentalize the emotional-semantic space within the whole lexico-semantic space. In the latter approach, in contrast, it would be reasonable to assume that emotion underlies the entirety of semantic space.

In addition, a theoretical position that affective information, regardless of its influence over memory performance, is not a part of semantics by definition would be conceivable; nevertheless, it must be noted that the above-mentioned position by Majerus and D'Argembeau (2011) recognizes an affective component of semantics. Furthermore, contemporary theories of semantic cognition suppose that semantic representations consist of various sources of information, such as sensory, motor, linguistic, and affective information (Lambon Ralph, Jefferies, Patterson, & Rogers, 2017; Martin, 2016). In particular, the neural and computational framework called controlled semantic cognition (CSC) (Lambon Ralph et al., 2017), which assumes two neural systems for semantic representation and semantic control, is relevant to the current study. According to the CSC framework, processes of semantic cognition are interpreted as controlled processes that manipulate semantic representations in a flexible way to generate appropriate behaviors within a task context. Even though we know various kinds of information about a concept, we have to select relevant information and ignore irrelevant information to realize appropriate behaviors (Jefferies & Lambon Ralph, 2006). Considering the commonality of affective dimensions in semantics is suggested by previous studies using a variety of words from different domains (Hollis & Westbury, 2016; Osgood & Suci, 1955), it is likely that the affective information of words (or of the concepts that words refer to), is typically selected and utilized to maintain words in STM. This interpretation would also explain why the SMS index based on the reduced number of three affective dimensions explains data from previous studies on STM better than the SMS_{w2v} index based on 300 abstract dimensions.

Limitations

The current study focuses on the semantic similarity effect on STM and implies that the component of semantic similarity by manipulations of previous studies affects STM. We, however, did not address nor quantify category-item or theme-item relationships because most of the previous studies reported some but not all categories or themes by which lists were constructed. It is beyond the scope of the current study, but it would be desirable to address the assumed cuing effect of category or theme labels directly in future studies (for the quantification of the label-item relationship, see Tse, 2009).

The creation of possible lists in our review was necessary due to the randomness inherent in previous studies' list construction methods, such as sampling words from a set or selecting words randomly from different sets; we had to infer possible lists for each study. Although we created the possible lists for each study based on original words that each study used, as a best attempt with available data, we acknowledge this approach to list creation based on inference as a limitation, which may have prevented calculation and reporting of

precise values for proposed indices. Our suggested list-level study, which will calculate and use these indices' values for each list, will overcome this limitation (see also the previous section regarding future directions).

Concluding remarks

In contrast to STM models' assumptions regarding and observations of similarity effects in STM such as the phonological similarity effect, semantic similarity appears to have a weak or null detrimental effect for serial reconstruction but a robust facilitative effect for serial recall by a common scoring of correct-in-position. In the current study, we have proposed a way of quantifying semantic similarity and an index for the strength of manipulation on similarity based on the affective dimensions of valence, arousal, and dominance. A review of data from previous studies along with our index suggests that semantic similarity has a detrimental effect on STM. Review on the semantic similarity effect based on the item versus order memory distinction further suggests that the detrimental effect of semantic similarity is on order memory.

Additionally, this study aimed to separate semantic similarity from semantic association both conceptually and statistically. Based on the review and meta-regression analysis on previous studies, it proposed that semantic similarity is detrimental to order memory while semantic association is facilitative to item memory. This differs from the common view that semantic similarity is detrimental (or neutral) to order memory while it is facilitative to item memory. Our proposed view would, by distinguishing semantic similarity from semantic association, aid in considering and theorizing semantic effects on STM in future studies.

Quantification of semantic similarity and estimation of manipulation strength on semantic similarity are important in research on the semantic similarity effect given that (a) observations of semantic similarity are likely dependent on manipulation strength arising from the selection of materials, (b) a manipulation does not necessarily work as intended, and (c) a manipulation might have effects above and beyond the similarity effect (e.g., association and a label's influence). Although defining psychological characteristics of semantics appears to be an intractable problem, the quantification of semantic similarity based on affective dimensions, as in the current study, or based on other types of dimensions, is a promising approach.

Acknowledgements This work was supported by a Grant-in-Aid for Japan Society for the Promotion of Science (JSPS) fellows [grant number 17J05372]. We thank Jean Saint-Aubin and two anonymous reviewers for their expert and insightful comments that improved our manuscript. We also thank Yusuke Takahashi for his advice on the meta-analysis of our work and Atsushi Tanimoto and Yuki Miyake for checking the mathematical equations.

Compliance with ethical standards

Conflicts of interest The authors report no conflicts of interest.

Open Practices All scripts (Python, R, and Shell scripts) used in this review are available at the first author's github repository (https://github.com/grocio/semantic-similarity-stm). This review was not preregistered.

References

References marked with an asterisk (*) indicate studies included in the meta-analysis or in the review with our proposed indices.

Avons, S. E., & Mason, A. (1999). Effects of visual similarity on serial report and item recognition. *The Quarterly Journal of Experimental Psychology Section A*, *52*(1), 217–240. doi:https://doi.org/10.1080/713755809

*Baddeley, A. D. (1966a). Short-term memory for word sequences as a function of acoustic, semantic and formal similarity. *Quarterly Journal of Experimental Psychology*, *18*(4), 362–365. doi:https://doi.org/10.1080/14640746608400055

*Baddeley, A. D. (1966b). The influence of acoustic and semantic similarity on long-term memory for word sequences. *Quarterly Journal of Experimental Psychology*, *18*(4), 302–309. doi:https://doi.org/10.1080/14640746608400047

Baddeley, A. D. (1986). Working memory. Oxford, UK: Oxford University Press.

Baddeley, A. D., & Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation* (Vol. 8, pp. 47–89). doi:https://doi.org/10.1016/S0079-7421(08)60452-1

Baddeley, A. D., Lewis, V., & Vallar, G. (1984). Exploring the articulatory loop. *The Quarterly Journal of Experimental Psychology Section A*, *36*(2), 233–252. doi:https://doi.org/10.1080/14640748408402157

*Belleville, S., Caza, N., & Peretz, I. (2003). A neuropsychological argument for a processing view of memory. *Journal of Memory and Language*, *48*(4), 686–703. doi:https://doi.org/10.1016/S0749-596X(02)00532-6

Bestgen, Y., & Vincze, N. (2012). Checking and bootstrapping lexical norms by means of word similarity indexes. *Behavior Research Methods*, *44*(4), 998–1006. doi:https://doi.org/10.3758/s13428-012-0195-z

*Biegler, K. A. (2007). Competition and inhibition in lexical retrieval: Are common mechanisms used in language and memory tasks? [Ph.D., Rice University]. In *ProQuest Dissertations and Theses* (304817719). ProQuest Dissertations & Theses A&I. https://search.proquest.com/docview/304817719?accountid=11929

Bourassa, D. C., & Besner, D. (1994). Beyond the articulatory loop: A semantic contribution to serial order recall of subspan lists. *Psychonomic Bulletin & Review*, *1*(1), 122–125. doi:https://doi.org/10.3758/BF03200768

Bradley, M. M., & Lang, P. J. (1999). *Affective Norms for English Words (ANEW): Instruction manual and affective ratings.* Technical Report C-1, The Center for Research in Psychophysiology, University of Florida.

Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*(3), 539–576. doi:https://doi.org/10.1037/0033-295X.114.3.539

Campoy, G., Castellà, J., Provencio, V., Hitch, G. J., & Baddeley, A. D. (2015). Automatic semantic encoding in verbal short-term memory: Evidence from the concreteness effect. *Quarterly Journal of*

Experimental Psychology, *68*(4), 759–778. doi:https://doi.org/10.1080/17470218.2014.966248

*Chassé, V., & Belleville, S. (2009). Input and output modes modulate phonological and semantic contributions to immediate serial recall: Evidence from a brain-damaged patient. *Cognitive Neuropsychology*, *26*(2), 195–216. doi:https://doi.org/10.1080/02643290902868534

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428. doi:https://doi.org/10.1037/0033-295X.82.6.407

Conrad, R. (1964). Acoustic confusions in immediate memory. *British Journal of Psychology*, *55*(1), 75–84. doi:https://doi.org/10.1111/j.2044-8295.1964.tb00899.x

*Crowder, R. G. (1979). Similarity and order in memory. In *Psychology of Learning and Motivation* (Vol. 13, pp. 319–353). doi:https://doi.org/10.1016/S0079-7421(08)60086-9

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods*, *51*(3), 987–1006. doi:https://doi.org/10.3758/s13428-018-1115-7

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, *315*(7109), 629–634. doi:https://doi.org/10.1136/bmj.315.7109.629

Farrell, S. (2006). Mixed-list phonological similarity effects in delayed serial recall. *Journal of Memory and Language*, *55*(4), 587–600. doi:https://doi.org/10.1016/j.jml.2006.06.002

Google. (2013). Project page for word2vec. Retrieved from https://code.google.com/archive/p/word2vec/

*Guérard, K., & Saint-Aubin, J. (2012). Assessing the effect of lexical variables in backward recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(2), 312–324. doi:https://doi.org/10.1037/a0025481

Gupta, P., Lipinski, J., & Aktunc, E. (2005). Reexamining the phonological similarity effect in immediate serial recall: The roles of type of similarity, category cuing, and item recall. *Memory & Cognition*, *33*(6), 1001–1016. doi:https://doi.org/10.3758/BF03193208

Haberlandt, K., Thomas, J. G., Lawrence, H., & Krohn, T. (2005). Transposition asymmetry in immediate serial recall. *Memory*, *13*(3–4), 274–282. doi:https://doi.org/10.1080/09658210344000297

*Hadley, C. B. (2006). Long-term memory contributions to verbal short - term memory performance in young and older adults [Ph.D., University of California, Los Angeles]. In *ProQuest Dissertations and Theses* (305340378). Health & Medical Collection; ProQuest Dissertations & Theses A&I. https://search.proquest.com/docview/305340378?accountid=11929

Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, D. D. (2019). Doing meta-analysis in R: A hands-on guide. Retrieved from https://bookdown.org/MathiasHarrer/Doing_Meta_Analysis_in_R/

Henley, N. M. (1969). A psychological study of the semantics of animal terms. *Journal of Verbal Learning and Verbal Behavior*, *8*(2), 176–184. doi:https://doi.org/10.1016/S0022-5371(69)80058-7

Hill, F., Reichart, R., & Korhonen, A. (2015). SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665–695. doi:https://doi.org/10.1162/COLI_a_00237

Hollis, G., & Westbury, C. (2016). The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic Bulletin & Review*, *23*(6), 1744–1756. doi:https://doi.org/10.3758/s13423-016-1053-2

Hulme, C. (2003). High- and low-frequency words are recalled equally well in alternating lists: Evidence for associative effects in serial recall. *Journal of Memory and Language*, *49*(4), 500–518. doi:https://doi.org/10.1016/S0749-596X(03)00096-2

Hulme, C., Maughan, S., & Brown, G. D. A. (1991). Memory for familiar and unfamiliar words: Evidence for a long-term memory contribution to short-term memory span. *Journal of Memory and Language*, *30*(6), 685–701. doi: https://doi.org/10.1016/0749-596X(91)90032-F

Hulme, C., Roodenrys, S., Schweickert, R., Brown, G. D. A., Martin, S., & Stuart, G. (1997). Word-frequency effects on short-term memory tasks: Evidence for a redintegration process in immediate serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(5), 1217–1232. doi: https://doi.org/10.1037/0278-7393.23.5.1217

Hurlstone, M. J., Hitch, G. J., & Baddeley, A. D. (2014). Memory for serial order across domains: An overview of the literature and directions for future research. *Psychological Bulletin*, *140*(2), 339–373. doi:https://doi.org/10.1037/a0034221

Hutchison, K. A. (2003). Is semantic priming due to association strength or feature overlap? A microanalytic review. *Psychonomic Bulletin & Review*, *10*(4), 785–813. doi:https://doi.org/10.3758/BF03196544

Huttenlocher, J., & Newcombe, N. (1976). Semantic effects on ordered recall. *Journal of Verbal Learning and Verbal Behavior*, *15*(4), 387–399. doi:https://doi.org/10.1016/S0022-5371(76)90034-7

Ishiguro, S., & Saito, S. (2019). Social dimensions in similarity judgment of faces. *Psychologia*. *61*(4), 252-268. https://doi.org/10.2117/psysoc.2019-A114

Jefferies, E., Hoffman, P., Jones, R., & Lambon Ralph, M. A. (2008). The impact of semantic impairment on verbal short-term memory in stroke aphasia and semantic dementia: A comparative study. *Journal of Memory and Language*, *58*(1), 66–87. doi:https://doi.org/10.1016/j.jml.2007.06.004

Jefferies, E., & Lambon Ralph, M. A. (2006). Semantic impairment in stroke aphasia versus semantic dementia: A case-series comparison. *Brain*, *129*(8), 2132–2147. doi:https://doi.org/10.1093/brain/awl153

Kousta, S.-T., Vinson, D. P., & Vigliocco, G. (2009). Emotion words, regardless of polarity, have a processing advantage over neutral words. *Cognition*, *112*(3), 473–481. doi:https://doi.org/10.1016/j.cognition.2009.06.007

Kowialiewski, B., & Majerus, S. (2018). The non-strategic nature of linguistic long-term memory effects in verbal short-term memory. *Journal of Memory and Language*, *101*, 64–83. doi:https://doi.org/10.1016/j.jml.2018.03.005

Kuperman, V., Estes, Z., Brysbaert, M., & Warriner, A. B. (2014). Emotion and language: Valence and arousal affect word recognition. *Journal of Experimental Psychology: General*, *143*(3), 1065–1081. doi:https://doi.org/10.1037/a0035669

Lambon Ralph, M., Jefferies, E., Patterson, K., & Rogers, T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, *18*(1), 42–55. doi:https://doi.org/10.1038/nrn.2016.150

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. doi:https://doi.org/10.1037/0033-295X.104.2.211

Logie, R. H., Brockmole, J. R., & Jaswal, S. (2011). Feature binding in visual short-term memory is unaffected by task-irrelevant changes of location, shape, and color. *Memory & Cognition*, *39*(1), 24–36. doi:https://doi.org/10.3758/s13421-010-0001-z

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203–208. doi:https://doi.org/10.3758/BF03204766

Ma, S., Zhang, Y., Liu, N., Xiao, W., Li, S., Zhang, G., … Ye, Z. (2019). Altered transposition asymmetry in serial ordering in early Parkinson's disease. *Parkinsonism & Related Disorders*, *62*, 62–67. doi:https://doi.org/10.1016/j.parkreldis.2019.01.028

Majerus, S., & D'Argembeau, A. (2011). Verbal short-term memory reflects the organization of long-term memory: Further evidence from short-term memory for emotional words. *Journal of Memory*

and Language, *64*(2), 181–197. doi: https://doi.org/10.1016/j.jml.2010.10.003

Martin, A. (2016). GRAPES—Grounding representations in action, perception, and emotion systems: How object properties and categories are represented in the human brain. *Psychonomic Bulletin & Review*, *23*(4), 979–990. doi:https://doi.org/10.3758/s13423-015-0842-3

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience*, *4*(4), 310–322. doi:https://doi.org/10.1038/nrn1076

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*(4), 547–559. doi:https://doi.org/10.3758/BF03192726

Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament. *Current Psychology*, *14*(4), 261–292. doi:https://doi.org/10.1007/BF02686918

Mewhort, D. J. K., Shabahang, K. D., & Franklin, D. R. J. (2018). Release from PI: An analysis and a model. *Psychonomic Bulletin & Review, 25*(3), 932–950. https://doi.org/10.3758/s13423-017-1327-3

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. Retrieved from http://arxiv.org/abs/1301.3781

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, *3*(4), 235–244. doi: https://doi.org/10.1093/ijl/3.4.235

Miron, M. S. (1969). What is it that is being differentiated by the semantic differential? *Journal of Personality and Social Psychology*, *12*(3), 189–193. doi:https://doi.org/10.1037/h0027714

Monnier, C., & Syssau, A. (2008). Semantic contribution to verbal short-term memory: Are pleasant words easier to remember than neutral words in serial recall and serial recognition? *Memory & Cognition*, *36*(1), 35–42. doi: https://doi.org/10.3758/MC.36.1.35

Monnier, C., & Syssau, A. (2014). Affective norms for french words (FAN). *Behavior Research Methods*, *46*(4), 1128–1137. doi: https://doi.org/10.3758/s13428-013-0431-1

Murdock, B. B. (1976). Item and order information in short-term serial memory. *Journal of Experimental Psychology: General, 105*(2), 191–216. doi:https://doi.org/10.1037/0096-3445.105.2.191

Nairne, J. S. (1990). A feature model of immediate memory. *Memory & Cognition*, *18*(3), 251–269. doi:https://doi.org/10.3758/BF03213879

Neale, K., & Tehan, G. (2007). Age and redintegration in immediate memory and their relationship to task difficulty. *Memory & Cognition*, *35*(8), 1940–1953. doi:https://doi.org/10.3758/BF03192927

Nelson, D. L., Bennett, D. J., & Leibert, T. W. (1997). One step is not enough: Making better use of association norms to predict cued recall. *Memory & Cognition*, *25*(6), 785–796. doi: https://doi.org/10.3758/BF03211322

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402–407. doi:https://doi.org/10.3758/BF03195588

*Nelson, D. L., Reed, V. S., & McEvoy, C. L. (1977). Learning to order pictures and words: A model of sensory and semantic encoding. *Journal of Experimental Psychology: Human Learning & Memory*, *3*(5), 485–497. doi:https://doi.org/10.1037/0278-7393.3.5.485

Nimmo, L. M., & Roodenrys, S. (2004). Investigating the phonological similarity effect: Syllable structure and the position of common phonemes. *Journal of Memory and Language*, *50*(3), 245–258. doi: https://doi.org/10.1016/j.jml.2003.11.001

Nishiyama, R. (2014). Active maintenance of semantic representations. *Psychonomic Bulletin & Review*, *21*(6), 1583–1589. doi:https://doi.org/10.3758/s13423-014-0618-1

Osgood, C. E. (1952). The nature and measurement of meaning. *Psychological Bulletin*, *49*(3), 197–237. doi:https://doi.org/10.1037/h0055737

Osgood, C. E. (1969). On the whys and wherefores of E, P, and A. *Journal of Personality and Social Psychology*, *12*(3), 194–199. doi:https://doi.org/10.1037/h0027715

Osgood, C. E., & Suci, G. J. (1955). Factor analysis of meaning. *Journal of Experimental Psychology*, *50*(5), 325–338. doi:https://doi.org/10.1037/h0043965

Page, M. P. A., Madge, A., Cumming, N., & Norris, D. G. (2007). Speech errors and the phonological similarity effect in short-term memory: Evidence suggesting a common locus. *Journal of Memory and Language*, *56*(1), 49–64. doi: https://doi.org/10.1016/j.jml.2006.09.002

Page, M. P. A., & Norris, D. (1998). The primacy model: A new model of immediate serial recall. *Psychological Review*, *105*(4), 761–781. doi:https://doi.org/10.1037/0033-295X.105.4.761-781

*Poirier, M., & Saint-Aubin, J. (1995). Memory for related and unrelated words: Further evidence on the influence of semantic factors in immediate serial recall. *The Quarterly Journal of Experimental Psychology Section A*, *48*(2), 384–404. doi:https://doi.org/10.1080/14640749508401396

Poirier, M., & Saint-Aubin, J. (1996). Immediate serial recall, word frequency, item identity and item position. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *50*(4), 408–412. doi:https://doi.org/10.1037/1196-1961.50.4.408

Poirier, M., Saint-Aubin, J., Mair, A., Tehan, G., & Tolan, A. (2015). Order recall in verbal short-term memory: The role of semantic networks. *Memory & Cognition*, *43*(3), 489–499. doi:https://doi.org/10.3758/s13421-014-0470-6

R Core Team. (2019). R: A language and environment for statistical computing. Retrieved from doi:www.R-project.org

Recchia, G., & Louwerse, M. M. (2015). Reproducing affective norms with lexical co-occurrence statistics: Predicting valence, arousal, and dominance. *Quarterly Journal of Experimental Psychology*, *68*(8), 1584–1598. doi: https://doi.org/10.1080/17470218.2014.941296

Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50.

Rips, L. J., Shoben, E. J., & Smith, E. E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior*, *12*(1), 1–20. doi:https://doi.org/10.1016/S0022-5371(73)80056-8

Roediger, H. L., Watson, J. M., McDermott, K. B., & Gallo, D. A. (2001). Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin & Review*, *8*(3), 385–407. doi:https://doi.org/10.3758/BF03196177

Saint-Aubin, J., Guérard, K., Chamberland, C., & Malenfant, A. (2014). Delineating the contribution of long-term associations to immediate recall. *Memory*, *22*(4), 360–373. doi: https://doi.org/10.1080/09658211.2013.794242

Saint-Aubin, J., Ouellette, D., & Poirier, M. (2005). Semantic similarity and immediate serial recall: Is there an effect on all trials. *Psychonomic Bulletin & Review*, *12*(1), 171–177. doi:https://doi.org/10.3758/BF03196364

*Saint-Aubin, J., & Poirier, M. (1999a). Semantic similarity and immediate serial recall: Is there a detrimental effect on order information? *The Quarterly Journal of Experimental Psychology Section A*, *52*(2), 367–394. doi:https://doi.org/10.1080/713755814

Saint-Aubin, J., & Poirier, M. (1999b). The Influence of Long-term Memory Factors on Immediate Serial Recall: An Item and Order

Analysis. *International Journal of Psychology*, *34*(5–6), 347–352. doi: https://doi.org/10.1080/002075999399675

Saito, S., Logie, R. H., Morita, A., & Law, A. (2008). Visual and phonological similarity effects in verbal immediate serial recall: A test with kanji materials. *Journal of Memory and Language*, *59*(1), 1–17. doi: https://doi.org/10.1016/j.jml.2008.01.004

Schweickert, R. (1993). A multinomial processing tree model for degradation and redintegration in immediate recall. *Memory & Cognition*, *21*(2), 168–175. doi: https://doi.org/10.3758/BF03202729

Smyth, M. M., Hay, D. C., Hitch, G. J., & Horton, N. J. (2005). Serial position memory in the visual-spatial domain: Reconstructing sequences of unfamiliar faces. *The Quarterly Journal of Experimental Psychology Section A*, *58*(5), 909–930. doi:https://doi.org/10.1080/02724980443000412

Stadler, M. A., Roediger, H. L., & McDermott, K. B. (1999). Norms for word lists that create false memories. *Memory & Cognition*, *27*(3), 494–500. doi:https://doi.org/10.3758/BF03211543

Stuart, G., & Hulme, C. (2000). The effects of word co-occurance on short-term memory: Associative links in long-term memory affect short-term memory performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(3), 796–802. doi:https://doi.org/10.1037/0278-7393.26.3.796

Surprenant, A. M., Neath, I., & Brown, G. D. A. (2006). Modeling age-related differences in immediate memory using SIMPLE. *Journal of Memory and Language*, *55*(4), 572–586. doi: https://doi.org/10.1016/j.jml.2006.08.001

Talmi, D., Luk, B. T. C., McGarry, L. M., & Moscovitch, M. (2007). The contribution of relatedness and distinctiveness to emotionally-enhanced memory. *Journal of Memory and Language*, *56*(4), 555–574. doi: https://doi.org/10.1016/j.jml.2007.01.002

Talmi, D., & Moscovitch, M. (2004). Can semantic relatedness explain the enhancement of memory for emotional words? *Memory & Cognition*, *32*(5), 742–751. doi: https://doi.org/10.3758/BF03195864

Tehan, G. (2010). Associative relatedness enhances recall and produces false memories in immediate serial recall. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *64*(4), 266–272. doi:https://doi.org/10.1037/a0021375

*Tse, C.-S. (2009). The role of associative strength in the semantic relatedness effect on immediate serial recall. *Memory*, *17*(8), 874–891. doi:https://doi.org/10.1080/09658210903376250

Tse, C.-S. (2010). A negative semantic similarity effect on short-term order memory: Evidence from recency judgements. *Memory*, *18*(6), 638–656. doi:https://doi.org/10.1080/09658211.2010.499875

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*(1), 91–109. doi: https://doi.org/10.1348/000712608X318617

*Tse, C.-S., Li, Y., & Altarriba, J. (2011). The effect of semantic relatedness on immediate serial recall and serial recognition. *Quarterly Journal of Experimental Psychology*, *64*(12), 2425–2437. doi: https://doi.org/10.1080/17470218.2011.604787

Underwood, B. J., & Goad, D. (1951). Studies of distributed practice: I. The influence of intra-list similarity in serial learning. *Journal of Experimental Psychology*, *42*(2), 125–134. doi:https://doi.org/10.1037/h0061106

Viechtbauer, W. (2019). Package 'metafor'. Retrieved from https://cran.r-project.org/web/packages/metafor/index.html

Vigliocco, G., Vinson, D. P., Lewis, W., & Garrett, M. F. (2004). Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cognitive Psychology*, *48*(4), 422–488. doi:https://doi.org/10.1016/j.cogpsych.2003.09.001

Vinson, D. P., & Vigliocco, G. (2008). Semantic feature production norms for a large set of objects and events. *Behavior Research Methods*, *40*(1), 183–190. doi:https://doi.org/10.3758/BRM.40.1.183

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*(5), 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*(4), 1191–1207. doi:https://doi.org/10.3758/s13428-012-0314-x

Watkins, M. J., Watkins, O. C., & Crowder, R. G. (1974). The modality effect in free and serial recall as a function of phonological similarity. *Journal of Verbal Learning and Verbal Behavior*, *13*(4), 430–447. doi:https://doi.org/10.1016/S0022-5371(74)80021-6

Weeks, D. G. (1976). Semantic space and encoding space in short-term memory. *Bulletin of the Psychonomic Society*, *8*(5), 356–358. doi: https://doi.org/10.3758/BF03335165

Whiteman, H. L., Nairne, J. S., & Serra, M. (1994). Recognition and Recall-Like Processes in the Long-Term Reconstruction of Order. *Memory*, *2*(3), 275–294. doi: https://doi.org/10.1080/09658219408258949

Wickens, D. D., & Clark, S. (1968). Osgood dimensions as an encoding class in short-term memory. *Journal of Experimental Psychology*, *78*(4, Pt. 1), 580–584. doi:https://doi.org/10.1037/h0026643

Williamson, V. J., Baddeley, A. D., & Hitch, G. J. (2010). Musicians' and nonmusicians' short-term memory for verbal and musical sequences: Comparing phonological similarity and pitch proximity. *Memory & Cognition*, *38*(2), 163–175. doi:https://doi.org/10.3758/MC.38.2.163