Journal of Memory and Language xxx (2010) xxx–xxx

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Global similarity accounts of embedded-category designs: Tests of the Global Matching models

Angela M. Maguire [a,*], Michael S. Humphreys [a], Simon Dennis [b], Michael D. Lee [c]

[a] Key Centre for Human Factors and Applied Cognitive Psychology, University of Queensland, Australia
[b] Department of Psychology, University of Adelaide, Australia
[c] Department of Cognitive Sciences, University of California, Irvine, United States

ARTICLE INFO

Article history:
Received 2 November 2006
revision received 23 March 2010
Available online xxxx

Keywords:
Category length
False memory
Global Matching
Hypothesis testing
Bayesian analysis

ABSTRACT

This paper addresses two Global Matching predictions in embedded-category designs: the within-category choice advantage in forced-choice recognition (superior discrimination for test choices comprising a same-category distractor); and the category length effect in forced-choice and old/new recognition (a loss in discriminability with increases in the number of same-category list items). The old–new data is analyzed using a Bayesian approach (Dennis, Lee, & Kinnell, 2008), which evaluates the evidence for both the null and the alternative hypothesis. Across two experiments, no within-category choice advantage was observed for the associative or the taxonomic categories. A category length effect was observed for the associative categories in forced-choice recognition, but not for the taxonomic categories. Additionally, the Bayesian analysis indicated that only a minority of participants evidenced a category length effect in old/new recognition. Such findings question the theoretical underpinnings of the Global Matching models. Namely, that global similarity drives interference in recognition.

Item-noise models have enjoyed substantial success as parsimonious explanations for the global similarity effects observed in episodic recognition. Prominent in such explanations have been the Global Matching (GM) models (a particular class of item-noise model; e.g., CHARM: Eich, 1982; SAM: Gillund & Shiffrin, 1984; MINERVA2: Hintzman, 1988; TODAM: Murdock, 1982; Matrix Model: Pike, 1984), which produce the effects as a consequence of the memory-access mechanism used for retrieval. In contrast, context-noise models (e.g., Anderson & Bower, 1972; Dennis & Humphreys, 2001) have to postulate additional mechanisms or processes, which occur during either encoding or retrieval, to account for the impact that related list items have on recognition performance. As context-noise predictions are ambiguous with respect to global similarity effects, we focus instead on the predictions of the GM models.

The introduction of the GM models initiated the development of a series of psychological theories that relied on substantial levels of inter-word similarity; particularly for taxonomic category exemplars governed by class inclusion rules (e.g., Eich, 1985; Gillund & Shiffrin, 1984; Hintzman, 1988; Metcalfe, 1990). The models were primarily well-received because they appeared to resolve the issue of how the collective properties of the study items affected recognition performance for each test item. This global similarity approach to recognition was largely supported by extant research on list length effects employing unrelated stimuli; and subsequent research on global similarity effects employing related stimuli. However, a substantial body of evidence has since accumulated that questions the reliability of the list length effect (e.g., Dennis & Humphreys, 2001; Dennis, Lee, & Kinnell, 2008; but see

* Corresponding author. Address: Key Centre for Human Factors and Applied Cognitive Psychology, University of Queensland, St. Lucia Campus 4072, Brisbane, Australia.
E-mail address: angie@humanfactors.uq.edu.au (A.M. Maguire).

0749-596X/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2010.03.007

Cary & Reder, 2003 for an alternative interpretation). This series of findings motivated our examination of global similarity effects in the embedded-category design.

The embedded-category design

The study list of an embedded-category design comprises multiple sets of conceptual and/or perceptual categories. Such experiments typically manipulate the number of items in each list category (i.e., category length: long; short; one/none), and/or the strength of the relationship between the list items and the test items employed (i.e., assumed similarity: high; intermediate; low/unrelated). In old/new recognition procedures, participants are typically required to discriminate words that have been studied (targets) from related (same-category) distractors, and in some experiments, unrelated distractors. In forced-choice procedures, the test choices typically comprise either a same-category target and distractor (within-category choices) or a target and distractor from different list categories (between-category choices). Note that, by design, the distractors in both the within-category choices and between-category choices are similar to the list items, but are either similar or dissimilar to the concurrently presented target, respectively.

The majority of research employing the embedded-category design has favored the use of conceptual category sets of associatively related words. Nonetheless, some experiments employ taxonomically related words interchangeably (i.e., do not differentiate between the two types of materials). The associative categories are typically derived using free association norms (i.e., a cue word is provided and participants are instructed to produce the first related word that comes to mind, or as many related words as possible). The category generator is the normed cue (e.g., *sleep*), and the category instances are a subset of the normed responses to the cue (e.g., *bed*, *tired*, *dream*, *snooze,* and so on). The taxonomic categories are typically derived using controlled association norms (i.e., a taxonomic label is provided and participants are instructed to produce the first exemplar that comes to mind, or as many exemplars as possible). The category generator is the taxonomic label (e.g., *units of time*) and the category instances are a subset of the normed responses to the label (e.g., *hour*, *year*, *day*, *century*, and so on). Thus, taxonomic categories can be defined as a specific type of associative hierarchy, the basis of which is class inclusion (Bower, Clark, Lesgold, & Winzenz, 1969). Less frequently employed in embedded-category designs are perceptual category sets of linguistically related items (e.g., orthographic or phonemic categories) and geometrically related items (e.g., dot patterns and faces).

Global similarity effects

In their seminal review of the GM models; Clark and Gronlund (1996) stated that: "global similarity effects may turn out to provide the strongest evidence for global matching" (p. 57). This claim subsumes three GM predictions in embedded-category designs: (1) the within-category choice advantage in forced-choice recognition (superior discrimination for test choices comprising a same-category distractor); (2) the category length effect in forced-choice and old/new recognition (a loss in discriminability with increases in the number of same-category list items); and (3) the prototype effect in forced-choice and old/new recognition (a loss in discriminability with increases in the similarity between the list items and the test items employed). Due to similarities in matching assumptions, successor models such as REM (Shiffrin & Steyvers, 1997) make the same three predictions in embedded-category designs (given a sufficient level of inter-item similarity; Criss, 2006; Criss, personal communication). In fact, REM can be defined as a GM model using the general criteria specified by Humphreys, Pike, Bain, and Tehan (1989).

In this paper, we briefly discuss retrieval in the GM models, and the findings relevant to the efficacy of the models. We subsequently present the findings for the embedded-category design. As other authors have addressed violations of the prototype effect (Westerberg & Marsolek, 2003; see also Miller & Wolford, 1999), we focus instead on the category length effect and the within-category choice advantage. Furthermore, our investigation of the effects is limited to conceptually related words, as this was the primary focus of the psychological theories of inter-word similarity derived from the GM models (e.g., Eich, 1985; Gillund & Shiffrin, 1984; Hintzman, 1988; Metcalfe, 1990). We acknowledge, however, that a comprehensive treatment of the item-noise/context-noise controversy also requires an examination of recognition memory for faces (Criss & Shiffrin, 2004), scenic photographs (Tulving, 1981), and orthographically and phonemically related words (Criss, 2006). Nonetheless, before these results can be interpreted within an item-noise, context-noise, or hybrid framework, it is necessary to firmly establish the effect of conceptual similarity on word recognition.

GM models and relevant findings

In GM models, the context reinstated by the participant at test is used to activate the *set of items* that occurred during the study episode.[1] The set of activated items are matched in parallel to each test item and the degree of match indexes the strength of the test item in memory. A given class of test items will produce a distribution of strength values with a mean and variance that is a function of: (1) the number of items in the retrieval set and (2) the similarity between the items in the retrieval set and each test item. As the number of items in the retrieval set increases, and the similarity of those items to each test item increases, so does the mean and variance of the strength distribution for a particular class of test items.[2]

[1] Some GM models do not specify a role for context in the retrieval process. For example, earlier versions of TODAM (Murdock, 1982) did not use context, but did restrict the matching process to the list of items comprising the study episode. Consequently, such models produce performance that is essentially equivalent to that produced by models that use context to isolate list items from the other items in memory (e.g., SAM; Gillund & Shiffrin, 1984).

[2] SAM (Gillund & Shiffrin, 1984) does not logically require variance in the structure of the model, but the recognition model has always assumed that the variance of a match is proportional to the strength of the match between the test probe and the contents of episodic memory.

Please cite this article in press as: Maguire, A. M., et al. Global similarity accounts of embedded-category designs: Tests of the Global Matching models. *Journal of Memory and Language* (2010), doi:10.1016/j.jml.2010.03.007

In this manner, the strength retrieved from memory is affected both by the properties of each test item, and by the collective properties of the items presented during the study episode.

The majority of the GM predictions are driven by the following assumption: any experimental manipulation that increases the mean of a strength distribution will produce a concomitant increase in the variance of that distribution. As the increase in mean strength is typically equivalent for targets and distractors at a given level of the experimental manipulation; the majority of the predicted effects are driven by variance increases across the levels. For example, increases in list length leave the difference between the means of the associated target and distractor distributions unchanged, while the variance of the respective distributions increases. In GM models, this variance increase produces a greater overlap of the distributions and a loss in discriminability for long lists relative to short lists.

The finding of null (and negative) list strength effects (e.g., Murnane & Shiffrin, 1991; Ratcliff, Clark, & Shiffrin, 1990) created a dilemma for proponents of the GM models: the null list strength effect supported the rejection of the models, while reports of list length effects supported their preservation. To resolve the dilemma, GM theorists invoked the *differentiation hypothesis* (e.g., McClelland & Chappell, 1998; Shiffrin, Ratcliff, & Clark, 1990; Shiffrin & Steyvers, 1997). The hypothesis holds that an increase in the strength of a list item decreases its similarity to other list items, while increasing its association to the study context. More recently, the list length effect has been challenged.

Dennis and Humphreys (2001) argued that the list length effect is an artifact of a number of processes that operate in conjunction during recognition procedures to produce a loss in discriminability for long lists relative to short lists. Namely, differential retention intervals, lapses in attention, displaced item rehearsals, and failures to accurately reinstate the study context. Their findings have demonstrated that, under carefully controlled conditions, it is possible to diminish the impact of these processes; reducing list length effects to negligible levels. Cary and Reder (2003) employed several of the controls suggested by Dennis and Humphreys (2001) and, while they still observed significant list length effects, the size of the effects were dramatically reduced in the conditions that employed these controls. Most importantly, Dennis et al. (2008) observed reductions in list length effects when these controls were employed individually; and negligible list length effects when these controls were employed in combination. Such findings undermine the GM models, and successor models such as REM, which employ similar item-noise assumptions. That is, if the list length effect is an artifact, the empirical justification for the item-noise assumption is undermined, and the differentiation hypothesis is superfluous. In order to provide convergent evidence for this argument, we examine global similarity effects in embedded-category designs.

Ultimately, our concern with the GM models is not whether they should be accepted or rejected, but rather whether the conclusions that have been drawn from these models, and the associated research, are still valid (e.g., pervasive list length effects for 'unrelated' items and pervasive similarity effects for 'related' items) and should therefore play an ongoing role in model development.

Global similarity: The category length effect

The category length effect refers to the loss in discriminability that sometimes accompanies an increase in the number of same-category list items. Experiments examining the effect typically employ various levels of category length within subjects, and compare performance across category length. Several experiments have reported losses in discriminability with increases in category length (e.g., Arndt & Hirshman, 1998; Hintzman, 1988; Shiffrin, Huber, & Marinelli, 1995). GM models predict the effect because an increase in the number of same-category list items increases the mean and variance of the associated target and distractor distributions. At a given level of category length, the strength increment is equivalent for the distributions, so the distance between the means remains the same. Accordingly, the variance increase produces a greater overlap of the distributions, and losses in discriminability with increases in category length.

The majority of research examining the category length effect has been conducted using blocked associative category sets; with a large proportion of this data being contributed by experiments employing the Deese/Roediger–McDermott (DRM) paradigm (Deese, 1959; Roediger & McDermott, 1995). The DRM paradigm, a special instantiation of the embedded-category design, produces very high false alarm rates and robust category length effects (see also Shiffrin et al., 1995). However, research employing taxonomic categories in the standard embedded-category design has produced more variable data, with some researchers reporting large effects of category length (e.g., Dewhurst, 2001; Hintzman, 1988) and others reporting small effects (e.g., Tussing & Greene, 1999). A review by Neely and Tse (2009) also demonstrates large differences in the magnitude of the category length effect for taxonomic categories. Nonetheless, despite this variability, the fact that significant category length effects have been observed in both old–new and forced-choice procedures appears to strongly support their existence.

However, there are two issues with the aforementioned findings. The first issue is that the old–new procedure and the between-category choice procedure do not prevent category-level information from impacting the recognition decision. In an old–new test, the category-level information abstracted during study could influence the criterion set to accept a test item (see Benjamin & Bawa, 2004; Miller & Wolford, 1999). It is also possible that category-level and item-level information are combined prior to making a decision. For example, Murnane and Phelps (1993, 1995) have argued that the combined familiarity of the item and the background context can drive the decision process. In a forced-choice test, criterion setting is not a

Please cite this article in press as: Maguire, A. M., et al. Global similarity accounts of embedded-category designs: Tests of the Global Matching models. *Journal of Memory and Language* (2010), doi:10.1016/j.jml.2010.03.007

concern.[3] However, if the participant mistakenly assumes that category-level information is informative (rather than redundant) for between-category choices, this could reduce performance relative to that for within-category choices. Specifically, there is random variability in the familiarity of the old and new test items and in the familiarity of the list categories, even though an equal number of items from each category comprising the choice have been studied. For example, when choosing between two items, *carrot* and *dog*, the old item, *carrot*, may appear to be only slightly more familiar than the new item, *dog*. However, the category *animals* may be considerably more familiar than the category *vegetables*. In this instance, basing the decision on the familiarity of the category (or on the combined familiarity of the item and the category) will reduce recognition performance. Hintzman's (1988) within-category choice condition addresses this confound: for within-category choices, category-level information is equated because the old and new item are drawn from the same list category. Thus, the probability of a correct within-category choice provides a direct measure of item familiarity. Nonetheless, the category length effect in Hintzman's within-category choice condition was driven by one data point: there was no loss in discriminability from 1- to 3-items, and a moderate loss from 3- to 5-items. Thus, a within-subjects replication of the finding is warranted; particularly considering the relatively uncontrolled procedures Hintzman employed.[4]

The second issue with the category length findings in the extant literature is the variability observed in the effect when different conceptual categories (i.e., associative and taxonomic) are employed. One possible explanation for this variability is that associates simply share more similarity than do taxonomic exemplars. However, there are several alternative explanations for associative category length effects, which do not rely on Global Matching, that we will address in the "General discussion".

Global similarity: the within-category choice advantage

The within-category choice (similar distractor) advantage refers to the superior discriminability observed for forced-choice test alternatives comprising a same-category distractor. Experiments examining the effect typically manipulate choice type within subjects, and compare performance across choice type at a given level of category length. Several experiments have reported a similar distractor advantage in forced-choice recognition using distractors that are related to at least one list item (e.g., Dobbins, Kroll, & Liu, 1998; Hintzman, 1988; Tulving, 1981). GM models predict the effect because the models assume that the strengths of a same-category target and distractor are correlated: the correlated strength assumption. The correlation produces better discriminability by reducing the covariance (i.e., the variance in the *difference distribution*) for same-category test alternatives, relative to test alternatives comprising a distractor that is similar to an item in memory, but dissimilar to the target (see Clark & Gronlund, 1996; Hintzman, 1988).

Tulving (1981, Experiment 1) examined the within-category choice advantage using scenic photographs as stimuli. The photographs were split into left–right halves, one of which was studied (target) and one of which was not studied (related distractor). Choice type was manipulated within subjects and a within-category choice advantage was observed (M_{diff} = .06). Dobbins et al. (1998) replicated Tulving (1981, Experiment 1) and observed a somewhat smaller advantage for within-category choices (M_{diff} = .04). Hintzman (1988, Experiment 2) used taxonomically related words in an embedded-category design and manipulated choice type between subjects. The graphed results indicate the within-category choice advantage ranged from approximately .02–.06 ($M_{diff} \approx .04$) across category length (1-; 3-; and 5-items).

There are two issues with the aforementioned findings. First, Hintzman (1988, Experiment 2) conducted the only prior test of the correlated strength assumption using word stimuli in an embedded-category design, and choice type was manipulated between subjects. That is, Hintzman's (1988) within-category choice advantage could be explained by a between-group difference in test strategy. Namely, some participants in the between-category choice condition may have assumed that both category-level and item-level information were relevant to the discrimination, when in fact all categories were instantiated by the study list.

Second, no experiment has examined the within-category choice advantage using associative categories (i.e., Hintzman employed taxonomic categories). Thus, if high levels of inter-word similarity are responsible for the very high false alarm rates and the robust category length effects observed for associative categories (cf. the DRM paradigm), and if there is a reliable within-category choice advantage for less similar taxonomic exemplars; then associative categories should produce a robust within-category choice advantage.

In summary, our review of the extant literature indicated that it was necessary to examine responding to the two types of conceptual categories within subjects, using both old–new and forced-choice procedures (and a within-subject manipulation of choice type). We accord particular weight to the category length results for within-category choices, as this choice type eliminates the use of category-level information. That is, for within-category choices, category-level information fails to distinguish

[3] The assumption that criterion does not play a role in forced-choice recognition decisions holds to the extent that a uni-dimensional decision axis is employed by the participant: the multiple sources of information that contribute to an item's strength in memory must be combined to produce a single scalar value for each test item. The majority of recognition memory models (e.g., Murdock, 1982; Hintzman, 1984; Humphreys et al., 1989) assume that there a very large number of relevant sources of evidence, but that they are indeed combined in this manner to produce a single scalar value.

[4] Hintzman's (1988) 200-item study list was presented across a 4-page booklet. Participants were required to rate the words on an activity scale. The format of his study-test procedure provided no systematic control over the amount of exposure to individual items, to the lag between successive related items, or to retention intervals. This may have resulted in differential rehearsal for items across category length and/or disparities in retention intervals for items across category length. Furthermore, within a level of category length, some categories of items were possibly more salient than others due to the close presentation of successive items from the same category in some instances, and wide separation in others.

Please cite this article in press as: Maguire, A. M., et al. Global similarity accounts of embedded-category designs: Tests of the Global Matching models. *Journal of Memory and Language* (2010), doi:10.1016/j.jml.2010.03.007

between the alternatives, so it cannot be used as a basis for responding.

Experiments

We had two primary objectives in Experiments 1 and 2, and one secondary objective. The primary objectives concerned GM predictions in embedded-category designs; the secondary objective concerned organizational processes (e.g., categorization) in embedded-category designs. First, we were investigating the GM prediction of a within-category choice advantage in forced-choice recognition. Second, we were attempting to establish the generality of the category length effect. Third, we were concerned about the effect that categorization (e.g., same-category inter-item associations and/or item-category associations) may have on discrimination in embedded-category designs.

GM models assume that global similarity effects are produced by inter-word similarity. If this assumption is correct, we should observe a within-category choice advantage in forced-choice recognition and a category length effect in both old/new and forced-choice recognition. In both experiments, we employ a within-subject manipulation of choice type to determine whether the within-category choice advantage is observed within subjects. In Experiment 1, we employ a within-subject manipulation of category type (associative vs. taxonomic) to address the variability in findings for such stimuli in embedded-category designs. If the category length effect is larger for the associative categories, the within-category choice advantage should also be larger. To our knowledge, no research has examined performance for these two types of conceptual categories within subjects. In Experiment 2, we employed taxonomic categories exclusively, as category length effects were not observed for these stimuli in Experiment 1.

In experiments employing the embedded-category design, considerations of categorization effects (e.g., Bower et al., 1969; D'Agostino, 1969) have been conspicuously absent. This is extraordinary given the correspondence between the embedded-category design and procedures that have traditionally been used to investigate organizational processes in recall and recognition. Our concerns regarding the effect of categorization on performance motivated the inclusion of a series of control variables in our two experiments. In Experiment 1, we manipulated the separation between same-category list items (*separation type*: blocked vs. distributed), and the item tested from the 5-item categories (*item tested*: first presented vs. fifth presented). In Experiment 2, we retained the item-tested control, and introduced a test expectancy manipulation (*task order*: embedded-category recognition task prior to, or subsequent to, an unrelated-item recognition task). Our intention was to systematically control the design of the two experiments and analyze any consistent effects in the data. While the inclusion of these variables produced a very well-controlled design, none of the variables consistently affected performance. Thus, we report the data collapsed across the control variables, and do not include these variables in the analyses.

Analysis techniques

There has been increasing concern that the *d'* sensitivity estimate from the old–new procedure is not always adequate for differentiating changes in discriminability from changes in bias (Heathcote, Raymond, & Dunn, 2006; Ratcliff, Sheu, & Gronlund, 1992). There are similar concerns with the A' statistic (Benjamin, 2005). However, the data from the forced-choice procedure (the probability of a correct choice) provides a measure of item strength uncontaminated by criterion; and the probability of a correct within-category choice provides a measure of item strength that eliminates category-level information from the decision. Thus, we employed a forced-choice procedure as our primary index of sensitivity.

A weakness of the forced-choice procedure is that a null effect of category length could be produced by a failure to attend to the categorical structure of the study list. To eliminate this alternative explanation for our findings, we employed an old–new procedure using the same study-phase design. The probability of an *old* response is likely to be the most sensitive measure of the participants' propensity to attend to the categorical structure of the list. That is, it will reflect changes in both sensitivity and criterion, and changes in one or both of these constructs would provide an indication that participants were attending to the categorical structure of the list.

Although the old–new procedure was originally included to demonstrate that participants were sensitive to the categorical structure of the list, a Bayesian analysis of the old–new procedure became available (i.e., Dennis et al., 2008). The Bayesian analysis allows a test of the universality of the category length effect (i.e., the extent to which the findings hold for the vast majority of participants), which is an essential feature of the item-noise assumption employed by the GM models, and successor models, such as REM.

Method

Design

In Experiment 1, a $2 \times 2 \times 2 \times 2 \times 2$ mixed factorial design with one nested factor was employed. We manipulated the type of conceptual category (taxonomic vs. associative), the category length (5-item vs. 1-item), the item tested within the category (first item presented in the 5-item categories and the only item presented in the 1-item categories vs. fifth item presented in the 5-item categories and the only item presented in the 1-item categories), the category separation type (blocked vs. distributed), and the type of recognition test (old–new vs. forced-choice). In both the old–new and the two-alternative forced-choice procedures, the distractor item for the taxonomic categories was the exemplar with the highest associative connection to the label (i.e., most representative exemplar of the category); for the associative categories, the distractor item was the cue word used to generate the categories via free association (i.e., the prototype of the category). The type of recognition test, the separation type, and the item tested were between-subject variables; category type and category length were

Please cite this article in press as: Maguire, A. M., et al. Global similarity accounts of embedded-category designs: Tests of the Global Matching models. *Journal of Memory and Language* (2010), doi:10.1016/j.jml.2010.03.007

within-subject variables. Additionally, participants in the forced-choice test received both within-category choices and between-category choices (nested factor) in which the two test alternatives (target and distractor) were drawn from the same list category or different list categories, respectively. For ease of comprehension, the Experiment 1 design is depicted graphically in the upper and middle portion of Fig. 1. The experiment was run as the second task in a participant session following a recognition task that employed unrelated items (there was no overlap in the stimuli used in the two experiments).

The design of Experiment 2 was similar to the Experiment 1 design and, for ease of comprehension and comparison, is depicted graphically in the lower portion of Fig. 1. A $2 \times 2 \times 2 \times 2$ mixed factorial design with one nested factor was employed. We manipulated the task order (embedded-category task first vs. embedded-category task second), the category length (5-item vs. 1-item), the item tested within the category and list (first/only item presented in the 5- and 1-item categories in the first section of the list vs. fifth/only item presented in the 5- and 1-item categories in the fifth section of the list), and the type of recognition test (old–new vs. forced-choice). Task order, item tested, and type of recognition test were between-subject variables, and category length was a within-subjects variable. Additionally, participants in the forced-choice test received both within-category choices and between-category choices (nested factor) in which the two test alternatives (target and distractor) were drawn from either the same or from different list categories, respectively.

Participants

In Experiment 1, 216 students participated to fulfill a partial credit requirement of an introductory psychology course at the University of Queensland. In Experiment 2, 127 students participated (70% were recruited in the same manner as those that participated in Experiment 1; the remainder were recruited via the student union website and were paid AUD10 for an hour session). All participants spoke English as a first language and were assigned by order of appearance to replication blocks defined by the between-subject conditions in each experiment (i.e., all between-subject conditions were run concurrently).

Experiment 1: Blocked 24 Taxonomic Categories & 24 Associative Categories Category length: 12 * 1-item categories & 12 * 5-item categories			
forced-choice recognition: between + within		old-new recognition	
first/only item tested	fifth/only item tested	first/only item tested	fifth/only item tested
N = 37	N = 37	N = 13	N = 13
Experiment 1: Distributed 24 Taxonomic Categories & 24 Associative Categories Category length: 12 * 1-item categories & 12 * 5-item categories			
forced-choice recognition: between + within		old-new recognition	
first/only item tested	fifth/only item tested	first/only item tested	fifth/only item tested
N = 38	N = 38	N = 14	N = 13

Experiment 2: Distributed 48 Taxonomic Categories Category length: 24 * 1-item categories & 24 * 5-item categories							
forced-choice recognition: between + within				old-new recognition			
first/only item tested		fifth/only item tested		first/only item tested		fifth/only item tested	
1st N = 18	2nd N = 18	1st N = 18	2nd N = 18	1st N = 12	2nd N = 12	1st N = 12	2nd N = 12

Fig. 1. Design for Experiments 1 and 2.

Please cite this article in press as: Maguire, A. M., et al. Global similarity accounts of embedded-category designs: Tests of the Global Matching models. *Journal of Memory and Language* (2010), doi:10.1016/j.jml.2010.03.007

Participants were assigned to the old–new and forced-choice tests in a manner that roughly equated the number of data points contributing to each condition (i.e., No. of participants × No. of observations was approximately equal across conditions). In Experiment 1, seven participants were excluded because of equipment failure, and a further five participants were excluded for failure to follow instructions (based on observable behavior during the experiment: e.g., failing to attend to the screen during the study phase; recording items on a notepad during the study phase; answering mobile phone during experiment). In Experiment 2, one participant was excluded because of equipment failure, and a further six participants were excluded for failure to follow instructions (based on observable behavior during the experiment). All excluded participants were identified within their respective experimental sessions, their data was immediately discarded, and their experiment files were assigned to a participant in the following experimental session.

Materials

In Experiment 1, category sets of six words from each of 24 associative categories (Appendix A) and 24 taxonomic categories (Appendix B) were used as stimuli. The associative categories were adapted from Roediger and McDermott's (1995) lists. For each category of associates, the cue word (category generator), and five items with the highest associative connection to the cue were selected (for further details regarding item selection see Roediger & McDermott, 1995). The taxonomic categories were adapted from the Casey and Heath (1988) norms and the Battig and Montague (1969) norms.[5] A further two taxonomic categories were created by the authors. For the taxonomic categories, the six exemplars with the highest associative connection to the label (category generator) were selected, with the following restriction: each word was required to hold a singular category interpretation and semantic definition. That is, the exemplars had to be representative of the category to which they were assigned and that category only (e.g., '*minute*' was deleted from the '*time*' category because it can be interpreted as both a unit of time and a unit of size).

In Experiment 2, category sets of six words from 48 taxonomic categories (Appendix C) were adapted from Yoon et al. (2003). The six exemplars with the highest associative connection to the label, which met the selection restrictions applied in Experiment 1, were used as stimuli in the experiment. In both experiments, the taxonomic categories were selected to maximize within-category item similarity and minimize between-category item similarity.

[5] When selecting the taxonomic categories we were careful not to include any category of items that overlapped with the associative categories (e.g., "color" was rejected as a possible taxonomic category because "orange" would be a representative exemplar and "fruit" was one of the associative category prototypes), or with each other (we overlooked the inclusion of "car" as a "vehicle" and "makes of car" as a category, however as "car" always functioned as a distractor this would only contribute noise to the false alarm rate for the taxonomic categories). Finally, Australian norms (Casey & Heath, 1988) were used whenever possible when choosing between the two sources of exemplars.

Item selection

In Experiment 1, all participants studied a 144-item list of words containing 12 1-item categories and 12 5-item categories from each of 24 associative and 24 taxonomic categories. Assignment of the 24 categories within each type (associative and taxonomic) to a category length (1-item or 5-item) was random. The most representative exemplar in each taxonomic category and the prototype in each associative category were retained to be used as distractors at test. For the five remaining items in each category, a single item was randomly selected to be presented for the 1-item categories; for the 5-item categories, all five remaining items were presented. Words in the 5-item categories were randomly assigned to the list, such that they did not necessarily appear in their order of emission in response to the category generator (label: taxonomic category; prototype: associative category). Item selection was carried out in this manner for each participant.

In Experiment 2, all participants studied a 144-item list of words containing 24 1-item taxonomic categories and 24 5-item taxonomic categories. The distractor was randomly selected within each taxonomic category. All other details pertinent to item selection were the same as those employed in Experiment 1.

Study lists

In Experiment 1, approximately half of the participants received blocked presentation of the same-category items; for the remaining participants, the same-category items were distributed throughout the list. In the blocked lists, the presentation of the categories was randomized such that each category (1-item or 5-item) could occur equally often at any position throughout the list. In the distributed lists, each item in the 5-item categories was randomly assigned to a fifth of the list; for the 1-item categories, the assignment to the first or fifth section of the list was determined by the item-tested condition (presentation location of the target item). That is, in the first/only item-tested condition, the single item in the 1-item categories and the first item in the 5-item categories was assigned to the first fifth of the list; in the fifth/only item-tested condition, the single item in the 1-item categories and the fifth item in the 5-item categories was assigned to the last fifth of the list. Sections two to four of the study list contained the second, third, and fourth items from the 5-item categories. The order of presentation of items was randomized within each fifth of the list. A new study list was created for each participant.

In Experiment 2, half the participants received the embedded-category task first; the remaining participants received the embedded-category task subsequent to a recognition task employing unrelated items. There was no overlap in the stimuli used in the two tasks and the unrelated-item task took approximately 10 min to complete. In the embedded-category task, all participants received a distributed list, which was constructed in the same manner as the distributed list in Experiment 1.

Test lists

In Experiment 1, the 32-pair forced-choice tests comprised 16 within-category choices (four associative and four taxonomic at each level of category length: 1-item

Please cite this article in press as: Maguire, A. M., et al. Global similarity accounts of embedded-category designs: Tests of the Global Matching models. *Journal of Memory and Language* (2010), doi:10.1016/j.jml.2010.03.007

and 5-item), and 16 between-category choices (four associative and four taxonomic at each level of category length). The 96-item old–new tests comprised 48 targets and 48 distractors (12 associative and 12 taxonomic at each level of category length). In Experiment 2, the forced-choice test comprised 16 within-category choices (eight at each level of category length) and 16 between-category choices (eight at each level of category length). The old–new test comprised 48 targets and 48 distractors (24 at each level of category length). The order of presentation of the test choices/items was randomized and a new test list was created for each participant.

In the forced-choice tests, the target and distractor comprising each test pair were always selected from categories of the same type (Experiment 1: associative or taxonomic; Experiment 2: taxonomic only) and length (Experiments 1 and 2: 1-item and 5-item). Assignment of the items within a category length to choice type (between-category or within-category) was random. For the within-category choices, the target and distractor were drawn from the same list category, while between-category choices comprised a target and distractor from different list categories. Between-category choices were constructed randomly such that each category of items could be tested against any other category of items within a category type and length.

Procedure

The experiments were computer administered. Participants were instructed that they would receive a list of words; that the words would appear one at a time in the center of the screen for approximately 3 s; that they should read the words silently, and attend to all of the words carefully, as they their memory for the words would be tested later in the experiment. In Experiment 1, all words were presented in uppercase (24-point MS Sans Serif) so that proper nouns requiring capitalization were perceptually indistinguishable from the items that did not. In Experiment 2, all words were presented in lowercase (24-point MS Sans Serif). Each word in the study list remained in the center of the screen for 2800 ms, and was immediately replaced by the next list item. The 144 words took approximately 7 min to present.

In Experiment 1, all participants spent 5 min engaged in a visuo-spatial puzzle task prior to the test phase. This equated the average retention interval for the items tested in the blocked and distributed lists. In Experiment 2, all participants received distributed presentation of same-category items: participants in the first/only item-tested conditions spent 5 min in the puzzle task; those in the fifth/only item-tested conditions spent approximately 9.5 min in the puzzle task. This equated the retention interval for the items tested from the first and fifth section of the list. The puzzle was a 6×6 grid of patterned tiles, similar to a two-dimensional Rubix® cube. Participants were required to use a mouse to rearrange the tiles to restore an abstract geometric picture. Test instruction subsequently informed participants of the nature of the recognition test.

Participants in the forced-choice test conditions were informed that each test pair contained an old and a new item. They were instructed to indicate which member of the pair was *old* by clicking the item's button with the mouse. Participants in the old–new test conditions were instructed to indicate whether each singly presented item was *old* or *new* by clicking the buttons centered below the test item. The tests were self-paced: the two alternatives in the forced-choice tests, and each item in the old–new tests, remained on screen until a response was registered. Upon registration of the response, the next test pair/item was immediately presented. Participants were tested in groups of varying size (ranging from 1–5) in individual carrels.

Bayesian analysis

The Bayesian analysis of the old–new procedure (Dennis et al., 2008) contrasts an error-only model (in which *d'* differences between conditions are assumed to be a Gaussian with a mean of zero) with an error-plus-effect model (in which *d'* differences are assumed to be drawn from the sum of an error distribution and a positive effect distribution). We use diffuse priors, so that the data determine the result. Additionally, to avoid a commitment to a 'least substantive difference', we rely on the property that Bayesian inference automatically penalizes the more complex error-plus-effect model. The Bayesian analysis does not focus exclusively on the truth of the null hypothesis. Rather, it determines the likely rate with which participants are best modeled by the error-only model (null hypothesis) vs. the error-plus-effect model (alternative hypothesis).[6] The Dennis et al. (2008) paper provides a more detailed description of the method, and outlines a number of additional advantages with respect to the standard *d'* analysis. A Bayesian analysis of the forced-choice procedure is yet to be developed.

Results

Experiments 1 and 2 employed both forced-choice and old/new recognition procedures in an embedded-category design. Table 1 presents the probability of a correct choice in the forced-choice procedure. Table 2 presents the hit rate (HR) and false alarm rate (FAR) in the old–new procedure. The data for the two experiments is collapsed across the control variables (see Appendices D and E for the data broken down by these variables). The standard error of the mean is given in parentheses as a measure of variability, and to allow performance comparisons across conditions. An alpha level of .05 was adopted for all statistical analyses.

In the forced-choice tests, a series of 2×2 repeated measures ANOVAs were performed to examine the effect of choice type (between- vs. within-category) and category length (5-item vs. 1-item) on the probability of a correct choice (discrimination). In the old–new tests, a series of Bayesian analyses (Dennis et al., 2008) were performed to examine the effect of category length on the sensitivity estimate (*d'*). Additionally, in the old–new tests, a series of planned comparisons were performed to examine the

[6] The usual method for assessing which of two models is preferred in a Bayesian context is to calculate the Bayes' factor: the ratio of the probability of the data given Model 1, divided by the probability of the data given Model 2. Note that such a formulation is subject to the same objection that we raise regarding null hypothesis significance testing: a small proportion of participants can override the evidence of the majority. For this reason, we favor the rate formulation, as we believe it more accurately reflects the question of empirical interest.

Please cite this article in press as: Maguire, A. M., et al. Global similarity accounts of embedded-category designs: Tests of the Global Matching models. *Journal of Memory and Language* (2010), doi:10.1016/j.jml.2010.03.007

Table 1
Forced-choice recognition test: probability of a correct response for the associative and taxonomic categories as a function of choice type (between-category vs. within-category) and category length (1-item vs. 5-item) in Experiments 1 and 2.

Forced-choice recognition test		Taxonomic		Associative	
Experiment	Choice type	1 -item	5-item	1-item	5-item
Exp 1 (N = 150)					
	Between	.90 (.014)	.88 (.014)	.85 (.016)	.73 (.019)
	Within	.89 (.014)	.89 (.013)	.83 (.016)	.75 (.019)
Exp 2 (N = 72)					
	Between	.85 (.018)	.83 (.019)	–	–
	Within	.84 (.018)	.84 (.019)	–	–

Note. Bracketed figures denote standard errors.

Table 2
Old/new recognition test: hit rate (HR) and false alarm rate (FAR) for the associative and taxonomic categories as a function of category length (1-item vs. 5-item) in Experiments 1 and 2.

Old/new recognition test		Taxonomic		Associative	
Experiment	Measure	1-item	5-item	1-item	5-item
Exp 1 (N = 54)					
	HR	.79 (.019)	.83 (.020)	.70 (.025)	.75 (.025)
	FAR	.13 (.017)	.22 (.023)	.19 (.022)	.32 (.025)
Exp 2 (N = 48)					
	HR	.73 (.025)	.79 (.024)	–	–
	FAR	.17 (.022)	.27 (.024)	–	–

Note. Bracketed figures denote standard errors.

effect of category length on the probability of an *old* response. Category length and choice type (in the forced-choice tests) were within-subjects variables. In Experiment 1, recognition performance for the taxonomic and associative categories was analyzed separately in the old–new tests and the forced-choice tests. In Experiment 2, taxonomic categories were employed exclusively, and recognition performance was analyzed separately in the old–new tests and the forced-choice tests.

Forced-choice recognition: null hypothesis sensitivity analysis

The primary experimental variables in the forced-choice procedure were choice type and category length. First, no effect of choice type was observed for either the associative categories (Experiment 1 only; $F < 1$), or the taxonomic categories (Experiments 1 and 2; both $Fs < 1$), in the forced-choice tests. That is, there was no evidence for a within-category choice advantage in forced-choice recognition: combining the data for the two experiments and collapsing across category length and category type, the mean difference score and 95% confidence interval for choice type (within-category choice minus between-category choice) was .00 ± .018 ($SD = .134$; $N = 222$). As these comparisons were performed within conditions and within subjects, the null effect of choice type was independent of the category length manipulation, and the participants' encoding and test strategies, respectively.

Second, discrimination significantly decreased as category length increased for the associative categories (Experiment 1 only; $F(1, 149) = 37.32$, $MSE = .036$, $p < .0001$, $\eta_p^2 = .200$), but not for the taxonomic categories (Experiments 1 and 2; both $Fs < 1$), in the forced-choice tests. That is, for the taxonomic categories, there was no evidence of a category length effect in forced-choice recognition: combining the data for the two experiments, the mean difference score and 95% confidence interval for category length (1-item category minus 5-item category) was .01 ± .018 for the taxonomic categories ($SD = .135$; $N = 222$; Experiments 1 and 2). For the associative categories, the mean difference score and 95% confidence interval was .09 ± .030 ($SD = .189$; $N = 150$; Experiment 1 only). Nonetheless, the Bayesian analyses of the old–new sensitivity estimates (see "Bayesian Sensitivity Analysis") provide evidence that Global Matching is unlikely to have produced the significant effect of category length that was observed for the associative categories in the forced-choice procedure.

To supplement the null hypothesis significance testing of the forced-choice data, we report a power analysis for each of the category length comparisons, and for the choice type comparison collapsed across category type. In the extant literature there is no agreed upon 'least substantive difference' for the category length and choice type comparisons. Thus, we calculated effect sizes by substituting a range of mean difference scores (.01–.10) into the numerator of the equation, and employed the standard deviations from our own comparisons in the denominator (see values in the previous paragraph). An alpha level of .05 was adopted for the power analyses. The results of these calculations are graphed in Fig. 2, with the expected mean difference score on the *x*-axis and power on the *y*-axis. Note that the power values for the choice type comparison and the taxonomic category length comparison are superimposed in the graph due to the very high consistency in values (same *N*; similar *SD*; same *CI*).

In Hintzman's (1988, Experiment 2) graphed results, the mean difference scores for choice type (within-category choice minus between-category choice) ranged from approximately .02 (1-item categories) to .06 (5-item categories). Collapsing across category length, this locates the mean difference score somewhere in the vicinity of .04. Inspection of Fig. 2 demonstrates we had an extremely

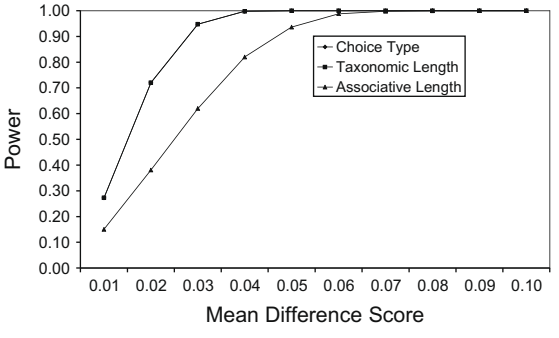

Fig. 2. Power analyses employing expected mean difference scores for the choice type comparison, and the taxonomic and associative category length comparisons.

Please cite this article in press as: Maguire, A. M., et al. Global similarity accounts of embedded-category designs: Tests of the Global Matching models. *Journal of Memory and Language* (2010), doi:10.1016/j.jml.2010.03.007

high chance of detecting a mean difference of this size (power ≈ 1.00). Thus, our findings strongly suggest that Hintzman's (1988, Experiment 2) between-subjects finding of a within-category choice advantage is not a robust effect. His finding may be explained by a between-group difference in test strategy (see "Introduction" for details).

Old/new recognition: encoding of category-level information

In order to establish that participants were encoding semantic information, we examined the effect of category length on the probability of an *old* response to targets and distractors in the old–new tests. In Experiment 1, the probability of an *old* response increased as category length increased for both the associative categories, $F(1, 52) = 20.28$, $MSE = .019$, $p < .0001$, $\eta_p^2 = .277$, and the taxonomic categories, $F(1, 52) = 17.59$, $MSE = .013$, $p = .0001$, $\eta_p^2 = .249$. Similarly, for the taxonomic categories in Experiment 2, the probability of an *old* response increased as category length increased, $F(1, 47) = 32.80$, $MSE = .008$, $p < .0001$, $\eta_p^2 = .411$. Such findings indicate that the participants were sensitive to the categorical structure of the list. That is, they were encoding semantic (category-level) information. Thus, the findings for the old–new test strongly suggest that the null effect of category length observed for the taxonomic categories in the forced-choice test is not due to a lack of semantic processing. Appendix F presents sensitivity (d') and bias (β) estimates for the old–new data, in addition to statistical analyses, to facilitate comparisons with published findings.

Bayesian sensitivity analysis

In the previous sections, we presented traditional null hypothesis significance tests and confidence intervals around the mean difference scores for the primary experimental variables (choice type and category length) in the forced-choice procedure. This procedure answers some important questions regarding sensitivity in embedded-category designs: the forced-choice data provides a measure of item strength uncontaminated by criterion; and the within-category choice data provides a measure of item strength uncontaminated by both criterion and category-level information. As we explained in the "Method" section of this paper, Dennis et al.'s (2008) Bayesian method for analyzing recognition memory can be applied to the old–new procedure. This Bayesian analysis of the old–new sensitivity estimate allows us to address a different question: how universal is the category length effect? Namely, is the effect observed for the vast majority of participants?

The Bayesian analysis was applied separately to the category length comparisons for: (1) the associative categories in Experiment 1; (2) the taxonomic categories in Experiment 1; and (3) the taxonomic categories in Experiment 2. Fig. 3 presents the posterior distributions of the rate with which participants are best modeled by the error-plus-effect model. Clearly, the probability mass is concentrated on the left for all three comparisons, indicating that most participants are best modeled by the error-only model.

As we have already discussed, Dennis et al.'s (2008) analysis produces a distribution of the rate with which participants should be assigned to the error-only model vs. the error-plus-effect model. To provide a concise statistic, we

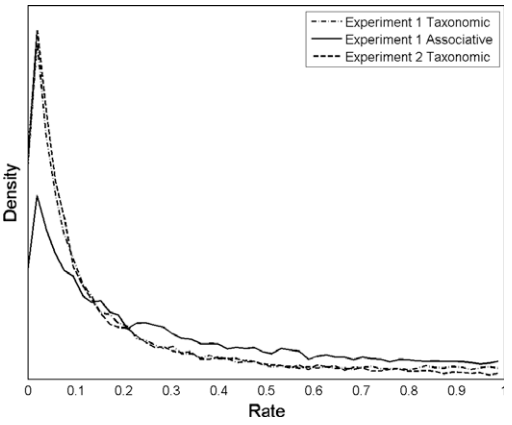

Fig. 3. The posterior distributions of the rate of selection of the error-plus-effect model for category length in the old–new test: the taxonomic categories in Experiment 1, the associative categories in Experiment 1, and the taxonomic categories in Experiment 2.

can address the question of how much of the probability mass falls below 10%, between 10% and 90%, and above 90%, given that these regions have equal prior probabilities (.33; .33; .33).[7] These three regions correspond to support for the error-only model, for both models, or for the error-plus-effect model, respectively. The three proportions add to 1.00 to summarize the full posterior distribution, indicating which model is most useful in accounting for the observed data.

In Experiment 1, the category length comparisons for both the associative categories (.74; .18; .08) and the taxonomic categories (.86; .09; .04) indicate that the probability that (at least) 90% of the participants are best captured by the error-only model is .74 for the associative categories, and .86 for the taxonomic categories, respectively. Similarly, in Experiment 2, the category length comparisons for the taxonomic categories (.89; .09; .02) indicate that the probability that (at least) 90% of the participants are best captured by the error-only model is .89.

In order to interpret the assigned proportions, we propose the following rule of thumb (c.f. Cohen's d; e.g., Cohen, 1992): when the probability of the assignment of 90% (or more) participants to a model is greater than .90, we have high confidence that it is a universal model; when the probability of the assignment of 90% (or more) participants to a model is between .90 and .70, we have moderate confidence that it is a universal model; when the probability is between .70 and .50, we have low confidence; and when it is between .50 and .33, we do not have sufficient evidence to draw a conclusion. Applying these criteria to the Bayesian analyses of

[7] Unlike null hypothesis significance testing, in which repeated analyses inflate the Type I error rate, the Bayesian analysis can be employed incrementally (Wagenmakers, 2006). As each participant is tested, one can run the analysis and stop when one of the following three outcomes becomes sufficiently probable: either positive evidence is obtained for the error-only model, for the error-plus-effect model, or for both (a mixed model). A high probability for the mixed model indicates that there is a reasonable proportion of participants that are performing as if there is no effect, and a reasonable proportion that are performing as if there is an effect. In this eventuality, one may have to consider redesigning the experiment, or at least consider why participants are not performing homogeneously.

Please cite this article in press as: Maguire, A. M., et al. Global similarity accounts of embedded-category designs: Tests of the Global Matching models. *Journal of Memory and Language* (2010), doi:10.1016/j.jml.2010.03.007

the old–new sensitivity estimates, we conclude that the error-only model (no category length effect) is favored as a universal model with moderate confidence for all three comparisons.

In the above analyses, we assumed an equal variance signal detection model (as most researchers do when employing the *d'* sensitivity estimate). However, there is a substantial literature suggesting that an unequal variance model is preferred (e.g., Ratcliff et al., 1992; Wixted, 2007). That is, in many situations, the ratio of the standard deviations of the new (noise) and old (signal) distributions is found to be approximately .80. Thus, we reanalyzed the data using this assumption, but found no substantive change in the conclusions (Experiment 1: associative categories [.75; .18; .07]; Experiment 1: taxonomic categories [.92; .07; .01]; Experiment 2: taxonomic categories [.72; .20; .08]).

More detailed results for the Bayesian analyses (equal variance model) are provided in Fig. 4: the upper panels present the posterior predictive distributions of the error-only and error-plus-effect models for all three comparisons; the lower panels present the modeled mean and 95% credible intervals for the observed differences in discriminability for each participant. As the graphs indicate, the observed data is much better accounted for by the error-only model, and the findings are highly consistent across participants.

In conclusion, the Bayesian analyses of the old–new data complement the findings for the power analyses of the forced-choice data. Additionally, the Bayesian analyses strongly suggest that the category length effect observed for the associative categories in the forced-choice test was not produced by the automatic global similarity process embodied by the GM models and successor models such as REM.

General discussion

In this paper we addressed two GM predictions in embedded-category designs: (1) the within-category choice advantage in forced-choice recognition (superior discrimination for test choices comprising a same-category distractor) and (2) the category length effect in forced-choice and old/new recognition (a loss in discriminability with increases in the number of same-category list items). Experiments demonstrating these effects have been cited (e.g., Clark & Gronlund, 1996; Hintzman, 1988) as providing the strongest support for the global similarity mechanism upon which the GM models are predicated. Nonetheless, our findings suggest that the effects are not robust.

Namely, there was no support for the within-category choice advantage in forced-choice recognition, and negligible support for the category length effect in forced-choice and old/new recognition. For the associative categories in the forced-choice test, we observed a category length effect, but no within-category choice advantage. For the taxonomic categories in the forced-choice-test, we observed no category length effect and no within-category choice advantage. For both types of conceptual categories in the old–new test, the Bayesian analyses of the sensitivity estimates indicated that category length effects were not observed for the vast majority of participants. Despite this failure to observe an automatic effect on sensitivity, the probability of an *old* response to both targets and distractors increased as category length increased – a direct indication that participants were processing semantic information during the experiment.

At this point it is clear that Clark and Gronlund (1996; see also Hintzman, 1988) were mistaken: for conceptually related words, global similarity effects are not robust, and do

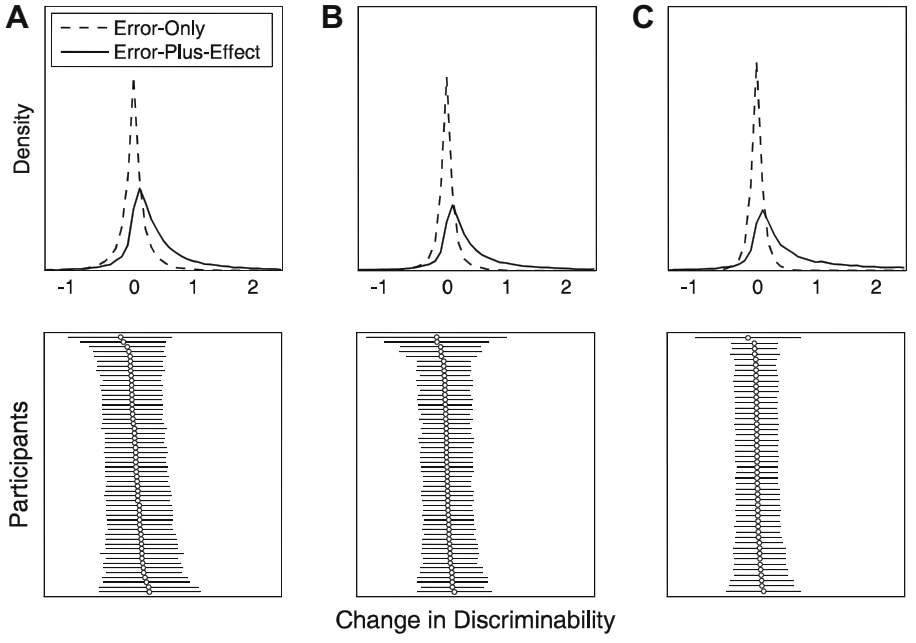

Fig. 4. Upper Panel: The posterior predictive distributions of the error-only and error-plus-effect models for the Experiment 1 associative categories (A), the Experiment 1 taxonomic categories (B), and the Experiment 2 taxonomic categories (C). Lower Panel: The modeled mean and 95% credible intervals for the observed differences in discriminability for each participant in each comparison.

Please cite this article in press as: Maguire, A. M., et al. Global similarity accounts of embedded-category designs: Tests of the Global Matching models. *Journal of Memory and Language* (2010), doi:10.1016/j.jml.2010.03.007

not provide strong support for the GM assumption. However, before concluding that our findings provide support for a context-noise account of recognition memory, it is necessary to explain why a minority of participants exhibited category length effects for the associative categories, and why category length effects for taxonomic categories are so variable when an old/new recognition test is employed.

Category length effects

The GM models provide a universal account of category length effects. However, the results of the Bayesian analyses suggest that a universal explanation for category length effects is unlikely. Rather, it seems possible that category length effects are multiply determined. In the following sections, we discuss a series of mechanisms and processes that may play differential roles in the production of category length effects. The roles of these mechanisms/processes may be partially determined by a participant's study and/or test strategy, and partially determined by the materials employed.

Implicit associative responses

Underwood's (1965) proposal, the implicit associative response (IAR), still appears to provide a viable explanation for some category length effects (when they are observed). In Underwood's (1965) conception, the IAR is a spontaneous process that produces symbolic information: single words or concepts that emerge into consciousness during the encoding of the study list (e.g., study: cat; IAR: dog). While Underwood assumes that participants are consciously aware of these associates, the responses are termed implicit because they are not overtly produced (Roediger & McDermott, 1995). Presumably, IARs are bound in memory with the study context during encoding, producing increases in the false recognition of such items at test. For example, in associative categories, all of the studied associates converge on the prototype (the category generator and the critical lure); whereas in taxonomic categories, they converge on the label (the category generator), not necessarily the most representative exemplar (the critical lure). This may explain some of the differences in performance for associative and taxonomic categories.

Cue substitution

It is also possible that recall processes (see Humphreys & Bain, 1983; Mandler, 1980) contribute to category length effects (when they are observed). Specifically, some similarity-based false alarms, and some reports of recollection in embedded-category designs (including the DRM paradigm), might derive from cue similarity, not target similarity. Cue substitution (a cue similarity effect) is a test phenomenon in which one cue acts as an alternate for another cue based on their associative or semantic similarity. False recall driven by cue substitution has been demonstrated in both semantic memory tasks (e.g., the Moses Illusion; Reder & Nesbit, 1991), and episodic memory tasks (e.g., Eich, 1982; Maguire, 2005).

Eich (1982, Experiment 4) provides a demonstration of cue substitution in cued recall. Her participants studied a list of unrelated cue-target pairs under instruction to form a meaningful inter-item association. At test, a small proportion of the intra-list cues were replaced with cue synonyms and target synonyms (extra-list cues related to one member of a study pair). Target recall to cue synonyms ($M = .29$) and cue recall to target synonyms ($M = .27$) was substantial. Recall intrusions that are produced by cue substitution introduce the possibility of a recognition analog.

Maguire (2005, Experiment 3) examined cue-substitution-driven false recall and false recognition within the same experimental design. Two groups of participants studied a list of related word pairs; at test, one group received a cued recall test and the other received a cue recognition test. In cued recall, a high rate of intra-list intrusions was observed for extra-list cues related to both members of a study pair (e.g., study: *bed-snooze*; test: *sleep*). As test instruction directed participants to retrieve using studied items only, it appeared that cue substitution was responsible for the false recall observed. When cue recognition was tested, the false alarm rate for distractors related to both members of a study pair was virtually identical to the intra-list intrusion rate in cued recall. Furthermore, an analysis comparing the derived recognition scores for recall participants with those for recognition participants confirmed that there was no difference in the probability of an *old* response for targets, related distractors, or unrelated distractors across the two test conditions. The procedural overlap and the consistency in performance for the two conditions supported the conclusion that the cue recognition participants were using the same strategy as the cued recall participants, and cue-substitution-driven recall-to-recognize was responsible for false alarms to related distractors in the recognition test.

The stimuli employed in Maguire (2005, Experiment 3) were not only constructed in the same manner as the associative categories used in Experiment 1 of this paper and Roediger and McDermott's (1995) study; the actual triads that were selected significantly overlapped with both sets of materials. The encoding of same-category inter-item associations is assumed to occur on a pair-wise basis in categorized lists (Neely & Balota, 1981). Thus, if same-category inter-item associations are encoded during the presentation of categorized lists in the same manner that inter-item associations are encoded between related pairs (as Neely & Balota, 1981, claim); Maguire's (2005, Experiment 3) findings bear significantly on the issue of whether recall affects recognition in embedded-category designs. That is, while Maguire's (2005) findings do not provide direct evidence for recall-to-recognize in embedded-category designs, they do suggest that a serious consideration of recall strategies is required.

Indeed, cue-substitution-driven recall-to-recognize appears to provide a legitimate test-trial alternative to Underwood's (1965) IAR account of false recognition in embedded-category designs. That is, rather than assuming that a strong associate of a list item is bound with the study context during encoding to create an episodic trace that is subsequently falsely recognized; a strong associate may act as a substitute cue for a stored episodic inter-item association between same-category items, producing the

Please cite this article in press as: Maguire, A. M., et al. Global similarity accounts of embedded-category designs: Tests of the Global Matching models. *Journal of Memory and Language* (2010), doi:10.1016/j.jml.2010.03.007

retrieval of a list item and providing the basis for the false recognition of a related distractor. Norman and Schacter's (1997, Experiment 2) research using the DRM paradigm is particularly relevant to this argument (see also Mather, Henkel, & Johnson, 1997). They had participants report the basis of their recognition responses along six-dimensions: memory for sound; list position; neighboring words; reactions; thoughts; and associations. They found that when participants reported basing their recognition response on a memory for a neighboring word (i.e., remembering a word that was presented either immediately before or after the test word) they could not discriminate between studied words and critical lures (i.e., the category prototype). The use of recall may also help to explain the inverted-U discrimination function for blocked taxonomic categories reported by Neely and Tse (2009).

Bias processes

It seems likely that category-level information may be involved in the production of category length effects (when they are observed). Miller and Wolford (1999) proposed a criterion-shift account of the category length effects observed in the DRM paradigm. However, their proposal has been discounted for several reasons. First, Miller and Wolford's (1999) data failed to differentiate their criterion-shift account (i.e., downward shifts in the criterion relative to static item distributions with increases in category length) from the standard sensitivity-based account (i.e., upward shifts in the item distributions relative to a static criterion with increases in category length; Wixted & Stretch, 2000). Second, direct attempts to manipulate criterion in the DRM paradigm (Gallo, Roediger, & McDermott, 2001) have failed to produce the predicted changes in the observed data.

Nonetheless, while Gallo et al. (2001) criticized Miller and Wolford's (1999) criterion-shift account; these authors also expressed support for the use of gist information in the recognition decision (Brainerd, Reyna, & Mojardin, 1999). Brainerd et al. (1999) assume that recognition decisions can be based either on a specific memory for a study event (verbatim trace), or on a memory for the gist of that event (fuzzy trace). The gist memory can be retrieved either when the study word is tested or when a related word is tested. The retrieval of the gist memory to either of these cues is assumed to produce a feeling of familiarity and a tendency to identify the test item as 'old'. A gist memory that 'matches' the studied and non-studied words from a taxonomic category (and tends to induce an 'old' response), may theoretically differ from a meta-knowledge that the members of a particular taxonomic category were in the study list (and the tendency to base an 'old' response on this meta-knowledge). Furthermore, it is also possible that there are phenomenological differences between a gist process and a meta-knowledge process. However, the two explanations are very difficult to empirically differentiate. Furthermore, using a multinomial model, it is easy to demonstrate that Miller and Wolford's (1999) criterion-shift account cannot be conceptually differentiated from the Brainerd et al. (1999) conception of a recognition decision based on gist. In fact, the two theories appear to be complementary in that the gist concept provides a mechanism for the abstraction of the category-level information necessary to produce a category-based criterion shift.

Finally, as Wixted and Stretch (2000) noted, there are alternative conceptions of a bias process that are not ruled out by their critiques of the Miller and Wolford (1999) criterion-shift account. One possibility is provided by Murnane and Phelps (1993, 1994, 1995). They obtained contextual reinstatement effects when words were studied in a unique combination of background color, screen location, and font color. That is, testing both targets and distractors in an old context, relative to a new context, increased the probability of an 'old' response for both targets and distractors. Their explanation was that the familiarity of the context added onto the familiarity of the item. If their proposal is correct then it is possible that category length effects may be produced by the familiarity of the category-level information adding onto the familiarity of the item-level information. Note that Higham and Brooks (1997) have demonstrated that participants can report the basic knowledge necessary to drive a criterion shift or other bias process (e.g., frequency, length, and grammatical class). Furthermore, Mulligan and Stone (1999) found it was necessary to have participants perform their experiments under divided attention to prevent category-level information from impacting performance on their implicit and explicit tasks.

Conclusions

We draw two conclusions from the results of our experiments. First, our data demonstrates that choice type and category length manipulations do not produce robust effects in well-controlled embedded-category designs employing conceptually related words. Consequently, proponents of GM models can no longer rely on these effects as providing strong and unique evidence for their theoretical position (see Clark & Gronlund, 1996; Hintzman, 1988). Second, our Bayesian analyses seriously question the universality of the category length effect and other semantic similarity effects that are reported in the literature (e.g., prototype effects). It is possible that similarity effects are near universal in the standard DRM paradigm, but not in simpler paradigms that better constrain encoding and test strategies. In summary, we shouldn't be focusing on universal explanations for recognition processes, which are typified by the GM models. Rather, we should be focusing on a variety of mechanisms and strategies that, in combination, produce conceptual similarity effects, including the large effects observed for recall and recognition in the DRM paradigm.

Acknowledgments

This paper is based on part of a dissertation submitted in fulfillment of the requirements for the PhD (Research) degree at the University of Queensland. This research was supported by a University of Queensland Graduate School Scholarship to A.M. Maguire, and by Grants DP0342656 and DP0556801 from the Australian Research Council to M.S. Humphreys.

Please cite this article in press as: Maguire, A. M., et al. Global similarity accounts of embedded-category designs: Tests of the Global Matching models. *Journal of Memory and Language* (2010), doi:10.1016/j.jml.2010.03.007

Appendix A. Experiment 1 stimuli: 24 associative categories

Adapted from Roediger and McDermott (1995).

ANGER	**GIRL**	**ROUGH**	**BLACK**	**HIGH**	**SLEEP**
MAD	BOY	SMOOTH	WHITE	LOW	BED
FEAR	DOLLS	BUMPY	DARK	CLOUDS	REST
HATE	FEMALE	ROAD	CAT	UP	AWAKE
RAGE	YOUNG	TOUGH	CHARRED	TALL	TIRED
TEMPER	DRESS	SANDPAPER	NIGHT	TOWER	DREAM
BREAD	**KING**	**SLOW**	**CHAIR**	**MAN**	**SOFT**
BUTTER	QUEEN	FAST	TABLE	WOMAN	HARD
FOOD	ENGLAND	LETHARGIC	SIT	HUSBAND	LIGHT
EAT	CROWN	STOP	LEGS	UNCLE	PILLOW
SANDWICH	PRINCE	LISTLESS	SEAT	LADY	PLUSH
RYE	GEORGE	SNAIL	COUCH	MOUSE	LOUD
COLD	**MOUNTAIN**	**SPIDER**	**DOCTOR**	**MUSIC**	**SWEET**
HOT	HILL	WEB	NURSE	NOTE	SOUR
SNOW	VALLEY	INSECT	SICK	SOUND	CANDY
WARM	CLIMB	BUG	LAWYER	PIANO	SUGAR
WINTER	SUMMIT	FRIGHT	MEDICINE	SING	BITTER
ICE	TOP	FLY	HEALTH	RADIO	GOOD
FOOT	**NEEDLE**	**THIEF**	**FRUIT**	**RIVER**	**WINDOW**
SHOE	THREAD	STEAL	APPLE	WATER	DOOR
HAND	PIN	ROBBER	VEGETABLE	STREAM	GLASS
TOE	EYE	CROOK	ORANGE	LAKE	PANE
KICK	SEWING	BURGLAR	KIWI	MISSISSIPPI	SHADE
SANDALS	SHARP	MONEY	CITRUS	BOAT	LEDGE

Note: Prototype presented in bold type.

Appendix B. Experiment 1 stimuli: 24 taxonomic categories

Adapted from Casey and Heath (1988) and Battig and Montague (1969).

HOUR	**BEER**	**DAISY**	**COPPER**	**HAMMER**	**DIAMOND**
YEAR	WHISKEY	CARNATION	STEEL	NAILS	RUBY
DAY	GIN	DAFFODIL	GOLD	SCREWDRIVER	SAPPHIRE
CENTURY	WINE	TULIP	ALUMINUM	CHISEL	OPAL
MONTH	VODKA	POPPY	SILVER	WRENCH	EMERALD
DECADE	BOURBON	PETUNIA	ZINC	PLIERS	PEARL
MAGAZINE	**MEASLES**	**CATHOLIC**	**TROUT**	**SPARROW**	**PINE**
BOOK	MUMPS	BUDDHIST	HERRING	BUDGIE	OAK
NEWSPAPER	LEPROSY	HINDU	SALMON	MAGPIE	EUCALYPTUS
PAMPHLET	MALARIA	METHODIST	TUNA	EAGLE	WILLOW
TEXTBOOK	SMALLPOX	BAPTIST	SNAPPER	COCKATOO	ELM
JOURNAL	HEPATITIS	MORMON	BARRAMUNDI	KOOKABURRA	MAPLE
CHURCH	**CHEMISTRY**	**COTTON**	**FORD**	**ANNE**	**CAR**
SYNAGOGUE	PHYSICS	WOOL	HOLDEN	MARY	TRUCK
TEMPLE	PSYCHOLOGY	SILK	TOYOTA	MARGARET	BUS
CHAPEL	BIOLOGY	RAYON	HONDA	JANE	TRAIN
CATHEDRAL	ZOOLOGY	NYLON	NISSAN	CATHY	MOTORBIKE
MOSQUE	BOTANY	LINEN	HYUNDAI	JENNY	PLANE
GUN	**JOHN**	**TENNIS**	**HOUSE**	**SYDNEY**	**OIL**
RIFLE	PETER	SOCCER	APARTMENT	PERTH	GAS
BOMB	DAVID	FOOTBALL	TENT	MELBOURNE	COAL
SWORD	PAUL	SWIMMING	HUT	ADELAIDE	WOOD

Please cite this article in press as: Maguire, A. M., et al. Global similarity accounts of embedded-category designs: Tests of the Global Matching models. *Journal of Memory and Language* (2010), doi:10.1016/j.jml.2010.03.007

Appendix B (*continued*)

PISTOL	ANDREW	CRICKET	HOTEL	BRISBANE	GASOLINE
CANNON	MICHAEL	HOCKEY	MOTEL	DARWIN	KEROSINE

Note: Most representative exemplar presented in bold type.

Appendix C. Experiment 2 stimuli: 48 taxonomic categories

Adapted from Yoon et al. (2003).

heart	ant	circle	king	fork	gun	doctor	morning
liver	bee	square	queen	knife	sword	teacher	noon
lungs	mosquito	triangle	prince	spoon	rifle	nurse	night
kidney	spider	rectangle	princess	spatula	spear	dentist	afternoon
stomach	ladybug	octagon	duke	ladle	bomb	lawyer	evening
brain	beetle	trapezoid	duchess	whisk	pistol	accountant	dawn
branch	piano	hammer	sergeant	milk	sparrow	happy	river
leaf	flute	nails	lieutenant	cheese	budgie	love	ocean
trunk	drum	screwdriver	captain	yogurt	magpie	sad	lake
root	saxophone	drill	colonel	ice-cream	eagle	hatred	stream
stem	trumpet	wrench	corporal	butter	cockatoo	anger	canal
twig	violin	pliers	admiral	cream	kookaburra	fear	sea
hour	biology	church	thunder	fairy	socks	rose	beer
day	chemistry	temple	rain	dragon	shirt	daisy	wine
year	physics	synagogue	hail	ghost	jumper	carnation	vodka
week	psychology	mosque	snow	witch	shoes	daffodil	rum
month	geology	cathedral	wind	mermaid	hat	tulip	gin
century	astronomy	chapel	lightning	goblin	dress	lily	whiskey
cow	magazine	refrigerator	car	necklace	blue	trout	house
pig	newspaper	stove	truck	ring	red	herring	apartment
horse	book	dishwasher	bus	bracelet	green	tuna	hut
chicken	pamphlet	microwave	bike	earring	yellow	salmon	shack
sheep	journal	dryer	train	watch	black	snapper	tent
goat	brochure	oven	plane	anklet	white	barramundi	hotel
nose	chair	brick	gas	basil	pine	potato	tennis
eyes	couch	wood	oil	oregano	oak	pumpkin	soccer
mouth	table	cement	diesel	thyme	eucalyptus	carrot	football
cheeks	bed	stone	coal	rosemary	willow	broccoli	swimming
ears	desk	steel	kerosene	parsley	elm	peas	cricket
lips	sofa	concrete	petroleum	garlic	maple	corn	hockey
oxygen	walnut	cotton	aunt	measles	apple	sydney	diamond
hydrogen	cashew	silk	uncle	mumps	banana	perth	ruby
nitrogen	almond	polyester	cousin	leprosy	pear	melbourne	sapphire
helium	pecan	wool	brother	malaria	peach	adelaide	opal
sodium	pistachio	rayon	sister	smallpox	grape	brisbane	emerald
potassium	macadamia	linen	mother	hepatitis	lemon	darwin	pearl

Appendix D. Old/new recognition data as a function of the control variables in Experiment 1 (separation type; item tested) and Experiment 2 (task order; item tested)

Forced-choice test		Taxonomic		Associative	
Condition	Choice type	1-item	5-item	1-item	5-item
Experiment 1: Blocked					
1st/only item presented ($N = 37$)	Between	.90 (.031)	.93 (.021)	.83 (.034)	.76 (.041)
	Within	.88 (.028)	.89 (.023)	.88 (.032)	.79 (.039)

(*continued on next page*)

Please cite this article in press as: Maguire, A. M., et al. Global similarity accounts of embedded-category designs: Tests of the Global Matching models. *Journal of Memory and Language* (2010), doi:10.1016/j.jml.2010.03.007

Appendix D (*continued*)

Forced-choice test		Taxonomic		Associative	
Condition	Choice type	1-item	5-item	1-item	5-item
5th/only item presented ($N = 37$)	Between	.91 (.024)	.85 (.030)	.87 (.028)	.72 (.036)
	Within	.89 (.028)	.89 (.025)	.80 (.037)	.74 (.036)
Experiment 1: Distributed					
1st/only item presented ($N = 38$)	Between	.89 (.028)	.84 (.033)	.84 (.033)	.70 (.038)
	Within	.89 (.026)	.89 (.029)	.80 (.030)	.71 (.044)
5th/only item presented ($N = 38$)	Between	.89 (.026)	.89 (.026)	.84 (.032)	.74 (.041)
	Within	.88 (.031)	.88 (.029)	.84 (.032)	.77 (.034)
Experiment 2: Cat task 1st					
1st/only item presented ($N = 16$)	Between	.85 (.034)	.80 (.042)	–	–
	Within	.82 (.049)	.82 (.027)	–	–
5th/only item presented ($N = 16$)	Between	.83 (.048)	.76 (.044)	–	–
	Within	.83 (.030)	.81 (.050)	–	–
Experiment 2: Cat task 2nd					
1st/only item presented ($N = 18$)	Between	.85 (.025)	.88 (.029)	–	–
	Within	.85 (.035)	.89 (.033)	–	–
5th/only item presented ($N = 17$)	Between	.87 (.036)	.87 (.033)	–	–
	Within	.85 (.029)	.86 (.036)	–	–

Appendix E. Old/new recognition data as a function of the control variables in Experiment 1 (separation type; item tested) and Experiment 2 (task order; item tested)

Old–new test		Taxonomic		Associative	
Condition	Measure	1-item	5-item	1-item	5-item
Experiment 1: Blocked					
1st/only item presented ($N = 13$)	HR	.72 (.040)	.84 (.041)	.68 (.040)	.79 (.028)
	FAR	.14 (.048)	.22 (.035)	.19 (.045)	.35 (.054)
5th/only item presented ($N = 13$)	HR	.76 (.042)	.79 (.053)	.67 (.058)	.62 (.054)
	FAR	.13 (.024)	.23 (.045)	.19 (.048)	.37 (.044)
Experiment 1: Distributed					
1st/only item presented ($N = 14$)	HR	.85 (034)	.80 (.027)	.69 (.049)	.79 (.038)
	FAR	.12 (.034)	.25 (.050)	.20 (.048)	.24 (.056)
5th/only item presented ($N = 14$)	HR	.82 (.034)	.88 (.032)	.76 (.053)	.78 (.057)
	FAR	.12 (.026)	.17 (.053)	.20 (.042)	.31 (.039)
Experiment 2: Cat task 1st					
1st/only item presented ($N = 12$)	HR	.74 (.049)	.80 (.050)	–	–
	FAR	.14 (.031)	.25 (.035)	–	–
5th/only item presented ($N = 12$)	HR	.63 (.060)	.68 (.054)	–	–
	FAR	.18 (.020)	.27 (.049)	–	–
Experiment 2: Cat task 2nd					
1st/only item presented ($N = 12$)	HR	.78 (.054)	.83 (.047)	–	–
	FAR	.19 (.055)	.31 (.058)	–	–
5th/only item presented ($N = 12$)	HR	.76 (.030)	.83 (.032)	–	–
	FAR	.18 (.047)	.24 (.049)	–	–

Please cite this article in press as: Maguire, A. M., et al. Global similarity accounts of embedded-category designs: Tests of the Global Matching models. *Journal of Memory and Language* (2010), doi:10.1016/j.jml.2010.03.007

Appendix F. Old/new recognition test: signal detection measures of discrimination (d') and bias (β) for the associative and taxonomic categories as a function of category length (1-item vs. 5-item) in Experiments 1 and 2

Old–new test		Taxonomic		Associative	
Separation Type	Measure	1-item	5-item	1-item	5-item
Experiment 1					
	d'	1.70 (.086)	1.54 (.083)	1.26 (.090)	1.01 (.091)
	β	1.35 (.078)	1.06 (.075)	1.31 (.077)	.99 (.055)
Experiment 2					
	d'	1.59 (.110)	1.44 (.096)	–	–
	β	1.57 (.145)	1.01 (.094)	–	–

Note: bracketed figures denote standard errors.

In Experiment 1, analyses of sensitivity (d') indicated that there was a significant effect of category length for the associative categories, $F(1, 52) = 5.96$, $MSE = .28$, $p = .02$,, $\eta_p^2 = .10$, but not the taxonomic categories, $F(1, 52) = 3.32$, $MSE = .19$, $p = .07$, $\eta_p^2 = .06$. In Experiment 2 (taxonomic categories only), category length also failed to significantly affect sensitivity, $F(1, 47)$, $MSE = .21$, $p = .13$, $\eta_p^2 = .05$.

In Experiment 1, analyses of bias (β) indicated that participants were more willing to respond *old* to items drawn from 5-item categories, relative to those drawn from 1-item categories, for both the associative categories, $F(1, 52) = 16.14$, $MSE = .17$, $p < .001$, $\eta_p^2 = .24$, and the taxonomic categories, $F(1, 52) = 9.36$, $MSE = .23$, $p = .004$, $\eta_p^2 = .15$. In Experiment 2 (taxonomic categories only), the analysis of bias also indicated a greater willingness to respond *old* to items from 5-item categories, $F(1, 47) = 18.35$, $MSE = .40$, $p < .001$, $\eta_p^2 = .28$.

References

Anderson, J. A., & Bower, G. H. (1972). Recognition and retrieval processes in free recall. *Psychological Review, 79*, 97–123.

Arndt, J., & Hirshman, E. (1998). True and false recognition in MINERVA2: Explanations from a global matching perspective. *Journal of Memory and Language, 39*, 371–391.

Battig, W. F., & Montague, W. E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology Monograph, 80*, 1–46.

Benjamin, A. S. (2005). Recognition memory and introspective remember/know judgments: Evidence for the influence of distractor plausibility on "remembering" and a caution about purportedly nonparametric measures. *Memory & Cognition, 33*, 261–269.

Benjamin, A. S., & Bawa, S. (2004). Distractor plausibility and criterion placement in recognition. *Journal of Memory and Language, 51*, 159–172.

Bower, G. H., Clark, M. C., Lesgold, A. M., & Winzenz, D. (1969). Hierarchical recall schemes in recall of categorized word lists. *Journal of Verbal Learning and Verbal Behavior, 8*, 501–506.

Brainerd, C. J., Reyna, V. F., & Mojardin, A. H. (1999). Conjoint recognition. *Psychological Review, 106*, 160–179.

Cary, M., & Reder, L. M. (2003). A dual-process account of the list-length and strength-based mirror effects in recognition. *Journal of Memory and Language, 49*, 231–248.

Casey, P. J., & Heath, R. A. (1988). Category norms for Australians. *Australian Journal of Psychology, 40*, 323–329.

Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin & Review, 3*, 37–60.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159.

Criss, A. (2006). The consequences of differentiation in episodic memory: Similarity and the strength based mirror effect. *Journal of Memory and Language, 55*, 461–478.

Criss, A. H., & Shiffrin, R. M. (2004). Context noise and item noise jointly determine recognition memory: A comment on Dennis & Humphreys (2001). *Psychological Review, 111*, 800–807.

D'Agostino, P. R. (1969). The blocked-random effect in recall and recognition. *Journal of Verbal Learning and Verbal Behavior, 8*, 815–820.

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

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic recognition memory. *Psychological Review, 108*, 452–478.

Dennis, S., Lee, M. D., & Kinnell, A. (2008). Bayesian analysis of recognition memory: The case of the list-length effect. *Journal of Memory and Language, 59*, 361–376.

Dewhurst, S. A. (2001). Category repetition and false recognition: Effects of instance frequency and category size. *Journal of Memory and Language, 44*, 153–167.

Dobbins, I. G., Kroll, N. E. A., & Liu, Q. (1998). Confidence-accuracy inversions in scene recognition: A remember-know analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 1306–1315.

Eich, J. M. (1982). A composite holographic associative recall model. *Psychological Review, 89*, 627–661.

Eich, J. M. (1985). Levels of processing, encoding specificity, elaboration, and CHARM. *Psychological Review, 92*, 1–38.

Gallo, D. A., Roediger, H. L., & McDermott, K. B. (2001). Associative false recognition occurs without strategic criterion shifts. *Psychonomic Bulletin & Review, 8*, 579–586.

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

Heathcote, A., Raymond, F., & Dunn, J. (2006). Recollection and familiarity in recognition memory: Evidence from ROC curves. *Journal of Memory and Language, 55*, 495–514.

Higham, P. A., & Brooks, L. R. (1997). Learning the experimenter's design: Tacit sensitivity to the structure of memory lists. *The Quarterly Journal of Experimental Psychology A, 50*, 199–215.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95*, 528–551.

Humphreys, M. S., & Bain, J. D. (1983). Recognition memory: A cue and information analysis. *Memory & Cognition, 11*, 583–600.

Humphreys, M. S., Pike, R., Bain, J. D., & Tehan, G. (1989). Global matching: A comparison of the SAM, MINERVA II, Matrix, and TODAM models. *Journal of Mathematical Psychology, 33*, 36–67.

Maguire, A. M. (2005). False alarms in episodic recognition: An examination of base-rate, similarity-based, and comprehensive theories. Unpublished PhD Thesis. Brisbane, Australia: University of Queensland.

Please cite this article in press as: Maguire, A. M., et al. Global similarity accounts of embedded-category designs: Tests of the Global Matching models. *Journal of Memory and Language* (2010), doi:10.1016/j.jml.2010.03.007

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review, 87*, 252–271.
Mather, M., Henkel, L. A., & Johnson, M. K. (1997). Evaluating characteristics of false memories: Remember/know judgments and memory characteristics questionnaire compared. *Memory & Cognition, 25*, 826–837.
McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105*, 734–760.
Metcalfe, J. (1990). Composite holographic associative recall model (CHARM) and blended memories in eyewitness testimony. *Journal of Experimental Psychology: General, 119*, 145–160.
Miller, M. B., & Wolford, G. L. (1999). Theoretical commentary: The role of criterion shift in false memory. *Psychological Review, 106*, 398–405.
Mulligan, N. W., & Stone, M. (1999). Attention and conceptual priming: Limits on the effects of divided attention in the category-exemplar production task. *Journal of Memory and Language, 41*, 253–280.
Murdock, B. B. (1982). A theory for the storage and retrieval of word and associative information. *Psychological Review, 89*, 609–626.
Murnane, K., & Shiffrin, R. M. (1991). Interference and the representation of events in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*, 855–874.
Murnane, K., & Phelps, M. P. (1993). A global activation approach to the effect of changes in environmental context on recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 882–894.
Murnane, K., & Phelps, M. P. (1994). When does a different environmental context make a difference in recognition? A global activation model. *Memory & Cognition, 22*, 584–590.
Murnane, K., & Phelps, M. P. (1995). Effects of changes in relative cue strength on context-dependent recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 158–172.
Neely, J. H., & Balota, D. A. (1981). Test expectancy and semantic organization effects in recall and recognition. *Memory & Cognition, 9*, 283–306.
Neely, J. H., & Tse, C. S. (2009). Category length produces an inverted – U discriminability function in recognition memory. *Quarterly Journal of Experimental Psychology, 62*, 1141–1172.
Norman, K. A., & Schacter, D. L. (1997). False recognition in younger and older adults: Exploring the characteristics of illusory memories. *Memory & Cognition, 25*, 838–848.
Pike, R. (1984). A comparison of convolution and matrix distributed memory systems. *Psychological Review, 91*, 281–294.
Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). The list-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 163–178.
Ratcliff, R., Sheu, C.-F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review, 99*, 518–535.
Reder, L. M., & Nesbit, G. W. (1991). Locus of the Moses Illusion: Imperfect encoding, retrieval, or match? *Journal of Memory and Language, 30*, 385–406.
Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 803–814.
Shiffrin, R., Huber, D., & Marinelli, K. (1995). Effects of category length and strength on familiarity in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 267–287.
Shiffrin, R. M., Ratcliff, R., & Clark, S. E. (1990). The list-strength effect: II. Theoretical mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 179–195.
Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM – Retrieving effectively from memory. *Psychonomic Bulletin & Review, 4*, 145–166.
Tulving, E. (1981). Similarity relations in recognition. *Journal of Verbal Learning and Verbal Behavior, 20*, 479–496.
Tussing, A. A., & Greene, R. L. (1999). Differential effects of repetition on true and false recognition. *Journal of Memory and Language, 40*, 520–533.
Underwood, B. J. (1965). False recognition produced by implicit verbal responses. *Journal of Experimental Psychology, 70*, 122–129.
Wagenmakers, E.-J. (2006). A practical solution to the pervasive problems of p-values. Accepted pending minor revisions. *Psychonomic Bulletin & Review, 1*, 1–10.
Westerberg, C. E., & Marsolek, C. J. (2003). Sensitivity reductions in false recognition: A measure of false memories with stronger theoretical implications. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 747–759.
Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114*, 152–176.
Wixted, J. T., & Stretch, V. (2000). The case against a criterion-shift account of false memory. *Psychological Review, 107*, 368–376.
Yoon, C., Feinberg, F., Hu, P., Gutchess, A. H., Hedden, T., Chen, H., et al. (2003). *Category norms as a function of culture and age: Comparisons of item responses to 105 categories by American and Chinese adults.* Unpublished manuscript, University of Michigan. <http://agingmind.beckman.uiuc.edu/Cat_Norms/index.html> Retrieved 01.09.04.

Please cite this article in press as: Maguire, A. M., et al. Global similarity accounts of embedded-category designs: Tests of the Global Matching models. *Journal of Memory and Language* (2010), doi:10.1016/j.jml.2010.03.007