Journal of Memory and Language 59 (2008) 361–376

Contents lists available at ScienceDirect

Journal of Memory and Language

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

Bayesian analysis of recognition memory: The case of the list-length effect

Simon Dennis [a,*], Michael D. Lee [b], Angela Kinnell [c]

[a] Department of Psychology, Ohio State University, 1835 Lazenby Hall, Columbus, OH 43201, USA
[b] Department of Cognitive Sciences, 3151 Social Sciences Plaza A, University of California, Irvine, CA 92697-5100, USA
[c] School of Psychology, North Terrace Campus, Level 4, Hughes Building, The University of Adelaide, SA 5005, Australia

ARTICLE INFO

Article history:
Received 17 October 2007
Revision received 20 June 2008
Available online 5 August 2008

Keywords:
Recognition memory
List-length effect
Bayesian methods
Context noise models
Item noise models
Word frequency effect
Signal detection theory

ABSTRACT

Recognition memory experiments are an important source of empirical constraints for theories of memory. Unfortunately, standard methods for analyzing recognition memory data have problems that are often severe enough to prevent clear answers being obtained. A key example is whether longer lists lead to poorer recognition performance. The presence or absence of such a list-length effect is a critical test of competing item- and context-noise based theories of interference and bares on whether recognition involves "recall-like" components as dual process theories would contend. However, the issue has remained unresolved, in part, because of the weaknesses of the standard analysis. In this paper, we develop a Bayesian method of analysis and apply it to new data on the list-length effect. The analysis allows us to find positive evidence in favor of a null list-length effect as predicted by context noise models. The data also illustrate the importance of the contextual reinstatement process on recognition performance and show how previous work demonstrating a list-length effect may have been contaminated by reinstatement confounds. By contrasting our new method against the standard approach we highlight the advantages of the Bayesian framework when inferring the values of psychologically meaningful variables, and in choosing between models representing different theoretical assumptions about memory.

In a typical yes/no recognition memory task, subjects are asked to study a list of items and then decide whether or not each of a set of test items appeared on the study list. This task has been a touchstone for understanding episodic memory (Glanzer & Adams, 1985; Ratcliff, Clark, & Shiffrin, 1990), and has provided important constraint for a series of memory models (Clark & Gronlund, 1996; Dennis & Humphreys, 2001; Eich, 1982; Gillund & Shiffrin, 1984; Hintzman, 1986; Humphreys, Bain, & Pike, 1989; McClelland & Chappell, 1998; Murdock, 1982; Shiffrin & Steyvers, 1997). Recently, however, there has been debate concerning the primary source of interference in recognition memory paradigms. Logically, interference can arise either from the other items that appear in the study list, or from the other contexts in which a test item has appeared, or from both sources (Humphreys, Wiles, & Dennis, 1994).

A critical empirical test of these competing theoretical positions involves the presence or absence of list-length effects. If item noise is the primary source of interference, recognition should be poorer for longer study lists than for shorter ones. If context is the primary source of interference, changes in the length of the study list should not change recognition performance. Currently, there is no consensus on whether or not a list-length effect is observed empirically, in part, because there are a number of confounds that could produce artifactual list-length effects.

The most obvious of these is the retention interval. If one presents a study list followed immediately by test, then retention interval will be longer for the long list. There are two ways in which retention interval can be

* Corresponding author.
E-mail address: simon.dennis@gmail.com (S. Dennis).

0749-596X/$ - see front matter
doi:10.1016/j.jml.2008.06.007

equated. In a retroactive condition, filler activity is added after the short list and only items from the start of the long list are tested. In a proactive condition, filler activity is added before the short list and only items from the end of the long list are tested.

Using the retroactive design, Schulman (1974) found no list-length effect in a forced choice test. Bowles and Glanzer (1983) did not analyze the retroactive condition separately from the proactive condition, but the difference in the proportion correct between short and long conditions was small (0.033) Also, in the third experiment of Murnane and Shiffrin (1991) the effect of length was not significant. In contrast to previous work, Gronlund and Elam (1994) did find a significant effect of length using a retroactive design. In this experiment, intentional instructions were employed and we will argue below that rehearsal could have been a factor.

In experiments employing proactive designs, the effects of length have been more robust. Bowles and Glanzer (1983) found a difference of 0.068 in the proactive condition, and overall found a significant effect of length. Underwood (1978) used a forced choice test and found an effect of length, as did Ohrt and Gronlund (1999). Underwood, citing the stability of word difficulty across list lengths and the lack of cumulative proactive interference in other recognition paradigms, argued against the direct involvement of proactive interference in recognition.

Rather, Underwood (1978) suggested that list-length effects in proactive designs were caused by a lack of attention. In long lists, subjects must maintain attention throughout the list. The items tested are those at the end of the list, which are the ones most likely to be affected by attentional lapses. In contrast, in short lists all items effectively appear at the start of the list. In the Bowles and Glanzer (1983) study, the long list contained 240 words. In the Underwood (1978) study, the long list contained 80 words and in the Ohrt and Gronlund (1999) study, the long list contained 82 words. In all three cases, words were presented for 1.5–2.0 s under intentional learning instructions, but with no specific processing requirements and no way of ensuring that attention was maintained. Lapses of attention seem likely under these conditions, particularly in the case of Ohrt and Gronlund (1999) in which subjects participated in four 50-min sessions.

A third potential confound is rehearsal. In the retroactive condition, a filler task is introduced between study and test. If subjects devote any of this time to rehearsing the studied items then performance in the short list will be superior to that in the long list both because there is more time to rehearse the short list and because any rehearsal that subjects might engage in under the long list conditions will be spread across more items and quite probably be focused on later items that will not be tested. Both experiments conducted by Gronlund and Elam (1994) involved intentional conditions, which increases the likelihood of rehearsal.

The fourth potential locus of an artifactual list-length effect, and the one on which we will focus in this paper, is contextual reinstatement. Episodic recognition necessarily involves the use of both an item and a context cue (Humphreys et al., 1994). In the retroactive design, subjects are either tested immediately in the long condition, or after the filler task in the short condition. After the long list, as far as the subject is aware, the current context can be used to initiate retrieval. However, in the case of the short list the current context focuses on the filler task, and so the subject is likely to reinstate the context of the study list so as to isolate the relevant study episode. To the extend that context drifts during the presentation of the long list, the end of list context may not be an efficacious cue for items that were presented at the start of the list and hence performance in the long list will suffer.

Controlling for the factors outlined above Dennis and Humphreys (2001) argued that, for verbal stimuli, context is the primary source of interference, and presented empirical evidence consistent with the absence of a list-length effect. Cary and Reder (2003) contested this conclusion, and presented empirical evidence consistent with a list length effect.

There were a number of differences between the two studies that could explain the different results. Cary and Reder (2003) only analyzed the combined proactive and retroactive results. As we argue above, list length effects in proactive designs have typically been larger than in retroactive designs, perhaps because of the effects of attention as suggested by Underwood (1978). Secondly, Cary and Reder (2003) employed the remember know procedure which requires subjects to attempt to recall specific aspects of the study episode. In recall, the existence of a list-length effect is not disputed, so it is possible that recall is contaminating the results in a way that did not occur in the Dennis and Humphreys (2001) experiments, which used yes/no recognition. Thirdly, Cary and Reder (2003) employed a much shorter period (2 min) between the end of the long list and test than did Dennis and Humphreys (2001, 8 min). It is possible that this shorter period was not sufficient to compel subjects to engage in contextual reinstatement in the long condition.

The source of interference is a fundamental aspect of understanding memory phenomena, and so this debate is crucial to the development of models of recognition memory. Unfortunately, the appropriate way to analyze recognition data has been a controversial topic (Banks, 1970; Lockhart & Murdock, 1970; Snodgrass & Corwin, 1988), because the methodology that is used standardly has a number of undesirable properties. These fall into two main classes: Those related to the application of signal detection theory (SDT), and those related to the application of standard methods for statistical inference. In this paper, we accept the standard SDT assumptions, but develop a Bayesian framework for understanding recognition memory performance that improves how the model can be related to experimental data. In particular, we tackle both issues of parameter estimation caused by the standard use of frequentist methods, and issues of model selection and evaluation caused by the standard use of null hypothesis significance testing (NHST).

We start by describing a new recognition memory experiment. We outline the method of analysis that would commonly be used in the recognition literature and, by applying it to the new data, describe its deficiencies. We then introduce and apply the Bayesian approach to the