Acta Psychologica 157 (2015) 155–163

Contents lists available at ScienceDirect

Acta Psychologica

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

Comparative difficulty and the strategic regulation of accuracy: The impact of test-list context on monitoring and meta-metacognition

Michelle M. Arnold *, Toby Prike

School of Psychology, Flinders University, Adelaide, SA, 5001, Australia

ARTICLE INFO

Article history:
Received 17 March 2014
Received in revised form 3 February 2015
Accepted 24 February 2015
Available online 22 March 2015

PsycINFO classification:
2340
2343
3040

Keywords:
Recognition
Context
Comparative difficulty
Strategic regulation
Confidence
Meta-metacognition

ABSTRACT

A growing body of research has shown that context manipulations can have little or no impact on accuracy performance, yet still significantly influence metacognitive performance. For example, participants in a test-list context paradigm study one list of words with a medium levels-of-processing task and a second word list with either a shallow or deep task: Recognition for medium words does not differ across conditions, however medium words are significantly more likely to be labeled as "remembered" (vs. merely familiar) if they had been studied with a shallow word list (Bodner & Lindsay, 2003). The goal of the current studies was to extend the test-list context paradigm to strategic regulation (report/withhold recognition test), and broaden it to incorporate different types of stimuli (i.e., face stimuli in place of a medium word list). The paradigm also was modified to include separate answer (studied/new) confidence and decision (report/withhold) confidence ratings at test. Results showed that context did not impact recognition accuracy for faces across the context conditions, however participants were more likely to report (i.e., volunteer) their face responses if they had studied the shallow word list. The results also demonstrated a difference between answer confidence and decision confidence, and the pattern of this difference depended on whether responses were reported or withheld (Experiment 1). Overall, the data are presented as support for the functional account of memory, which views memory states as inferential and attributional rather than static categories.

1. Introduction

Theories that have been used to explain the underlying mechanisms of recognition memory traditionally have fallen into one of two broad camps – quantitative versus qualitative accounts. Quantitative approaches typically argue that differences in recognition phenomenology (e.g., recollection vs. familiarity) are determined by decisional processes that are based on memory strength. For instance, some quantitative accounts postulate that the same memory traces feed into the different states of recollection and familiarity, but that factors such as trace strength contribute to whether the memory is experienced as one, the other, or both (Dunn, 2004, 2008; Heathcote, Bora, & Freeman, 2010). Conversely, qualitative approaches tend to focus on the fundamental principle that distinct underlying processes give rise to experiences of recollection and familiarity; that is, the two types of experiences are independent of one another and may occur in isolation or together (Gardiner, Kaminska, Dixon, & Java, 1996; Jacoby, 1991; Jacoby, Yonelinas, & Jennings, 1997).

More recently, though, there has been a shift to a new class of memory models that combines aspects of both qualitative and quantitative approaches. For example, the *functional account* characterizes both recollection and familiarity as inferential (see McCabe & Balota, 2007, for discussion of the similar *expectancy heuristic account*). Specifically, the context in which a memory decision is made will impact the outcome, and if the information that you have available about a specific event allows you to make the decision-at-hand (e.g., recognizing a person on the street as your psychology professor) then likely you will judge that you have conscious recollection. However, that same available information in a different context may only support a more general feeling of familiarity (e.g., recognizing that you know a person on the street, but not who they are; Arnold, 2011; Bodner & Lindsay, 2003; Gruppuso, Lindsay, & Kelley, 1997).

One key line of research that has lent support to the functional account of memory is Bodner and Lindsay's (2003) test-list context paradigm. This paradigm requires participants to study two separate lists of words; one list is always studied with a medium levels-of-processing (LOP) task, and the other is studied either with a shallow or deep LOP task. Specifically, for the words on the medium LOP study list participants make a yes/no judgment about whether the word is one people commonly use, whereas for the shallow LOP task they judge whether the word contains the letter 'a' and for the deep LOP task they judge whether the word is an item they would want on a deserted island. Participants then complete a remember-know (R-K) judgment recognition test, for which they are instructed that they

* Corresponding author at: School of Psychology, Flinders University, GPO Box 2100, Adelaide, SA, 5001, Australia. Tel.: +61 882 012 256; fax: +61 882 013 877.
E-mail address: michelle.arnold@flinders.edu.au (M.M. Arnold).

http://dx.doi.org/10.1016/j.actpsy.2015.02.018
0001-6918/

should use the "remember" (R) label if they are able to bring to mind specific details of having previously encountered a word, whereas the "know" (K) label is for words that do not have accompanying conscious details, but nonetheless feel "old" or familiar. Bodner and Lindsay found that accurate old/new discrimination did not differ for the medium items between the conditions, however participants in the shallow condition were significantly more likely to use the R label for their medium words (i.e., claim conscious recollection) than the participants in the deep condition.

Based on their findings, Bodner and Lindsay (2003) argued that it was the context at test that led participants in the shallow condition to use different attributes of the stimuli to define the properties of R and K than participants in the deep condition. For example, the context manipulation created a "comparative difficulty" difference between the two groups of participants; that is, the medium items were being recalled in either a context of harder-to-recall items (shallow list) or easier-to-recall items (deep list). Therefore, although the medium words produced similar objective memory patterns (old/new judgments), the difference in comparative difficulty between the shallow and deep context conditions led to significantly different subjective experience ratings (R vs. K judgments).

The main purpose of the current two experiments was to modify and extend the test-list context paradigm. The aforementioned test-list context experiments have always used the same type of stimuli across the medium and shallow/deep conditions (Bodner & Lindsay, 2003; Tousignant & Bodner, 2012; see also Bodner & Richardson-Champion, 2007, for an event-details test-list context paradigm). Therefore, one aim of the current experiments was to demonstrate that the comparative difficulty of one type of stimuli can influence performance on a *different type* of stimuli. As in previous test-list context experiments, the comparative difficulty variable was created by requiring participants to study either a shallow or deep word list, but the medium study list for all participants was a set of faces. Additionally, instead of using an R-K judgment (or recollection/familiarity ratings; Tousignant & Bodner, 2012) to assess the impact of comparative difficulty, performance was gauged within a strategic regulation paradigm. Participants were required to make an old/new recognition judgment for every test item, but they were allowed to choose whether to report each response for scoring (points/penalties for correct and incorrect responses, respectively) or to withhold if they were unsure about their response (no points/penalties). If the test-list context effect is not dependent on using one class of stimuli, or a performance measure that requires people to necessarily evaluate conscious recollective details, then the comparative difficulty of the word lists should influence the strategic regulation of accuracy for the face stimuli.

1.1. Comparative difficulty and metacognitive monitoring

Although R-K judgments routinely have been used to parse out memory processes, there are noted issues with this type of performance measure. For example, research has shown that one-step R-K judgments (no overt old/new judgment) promote more liberal responding than two-step judgments, and that the inclusion of a "guess" option improves accuracy discrimination (Eldridge, Sarfatti, & Knowlton, 2002; Gardiner, Ramponi, & Richardson-Klavehn, 2002; Tousignant, Bodner, & Arnold, 2015). Additionally, recent work with the independent Recollection-Familiarity ratings, which are argued to map conceptually to the R-K judgment (see Higham & Vokey, 2004, for more in-depth discussion), has shown that the ratings do not always produce the same pattern of results as an R-K judgment. For instance, Tousignant and Bodner (2012) showed that replacing the R-K judgment with independent rating scales in the test-list context paradigm produced both higher recollection and familiarity ratings for the medium items in the shallow condition. Finally, a potential weakness of R-K judgments (and Recollection-Familiarity ratings) is that they are not intuitive and/or natural for most individuals: Participants require intensive instructions and follow-up clarifications to comprehend R-K judgments, and usage of the judgments can vary substantially between-subjects (e.g., differences in willingness to label a detail as "conscious" recollection).

Given some of the potential issues of the R-K judgment, one goal of the current studies was to demonstrate a test-list context effect with a different performance measure; the strategic regulation of accuracy. Research on strategic regulation focuses on how we use the knowledge and understanding that we have of our cognitive processes and memories to direct our performance (e.g., deciding which items that come to mind should be volunteered on a free-recall memory test; Higham, 2007). Demonstrating a test-list context effect for strategic regulation not only would strengthen support for the functional account of memory, but it also would highlight the need to consider the more widespread influence of comparative difficulty. For example, comparative difficulty may impact strategic regulation on formula-scored exams, which require students to report answers they are confident in for points/penalties and withhold low-confidence/guess answers (0 points/penalties). It is possible that the same test question appearing on two different exams could yield different report/withhold outcomes that are due solely to differences in comparative difficulty between the tests. Specifically, students may be more likely to withhold their response to that question on an exam that contains easy questions because, by comparison, it seems more difficult to answer, whereas they may be more willing to report their response on a more difficult exam because in that situation the question feels comparatively easier to answer.

Two main approaches have been used to examine the strategic regulation of accuracy. Koriat & Goldsmith (1996; Goldsmith & Koriat, 2008) developed the first instrumental approach, which they refer to as the monitoring-control framework. This framework specifies that there are three crucial components that contribute to overall performance; retrieval, monitoring, and control. The main tenet is that, rather than just reporting every piece of information that comes to mind on a free recall task, there is a monitoring mechanism that subjectively evaluates the likelihood that each item is correct, and a control mechanism that determines whether the item should be reported (i.e., volunteered as correct information). Importantly, the control mechanism can be influenced by a variety of factors (e.g., the size of payoffs/penalties for reported information), which means an individual's decision to volunteer or withhold a particular piece of information may differ depending on the context in which s/he makes the decision (Higham, 2007; Koriat & Goldsmith, 1996). The second, more recent, approach comes from type-2 signal-detection theory (SDT) methodology, and it has predominantly been applied to strategic regulation of accuracy in both test-taking situations (Arnold, Higham, & Martín-Luengo, 2013; Higham, 2007; Higham & Arnold, 2007) and recognition memory (Arnold, 2013; Higham, Perfect, & Bruno, 2009). The type-2 SDT framework is similar in its constructs to the monitoring-control framework, but one of its main advantages is that the methodology permits for a straightforward and concrete measure of metacognitive monitoring. Therefore, with type-2 SDT it is possible to directly compare monitoring ability across different groups and conditions, and for this reason it is the framework adopted for the current paper (but see Higham, 2011, and Goldsmith, 2011, for an in-depth comparison of the two approaches).

Most memory researchers are familiar with type-1 SDT through its application in recognition memory to produce several useful measures of performance (e.g., Green & Swets, 1966; Rotello, Macmillan, & Reeder, 2004). For instance, researchers are able to calculate indices of accuracy, such as discrimination (how well you distinguish between signal and noise trials; e.g., d'), and response bias (how willing you are to claim a signal is present; e.g., C). Type-2 SDT uses the same general methodology/principles as type-1, but the key distinction is the nature of the underlying distributions. That is, type-1 SDT focuses on experimenter-defined distributions (e.g., studied vs. new items),