Journal of Memory and Language 90 (2016) 161–176

Contents lists available at ScienceDirect

Journal of Memory and Language

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

Recalibration effects in judgments of learning: A signal detection analysis ☆

Katarzyna Zawadzka [a,b,*], Philip A. Higham [a]

[a] Psychology, University of Southampton, UK
[b] School of Psychology, Cardiff University, UK

ARTICLE INFO

Article history:
Received 31 July 2015
revision received 12 April 2016

Keywords:
Judgments of learning
Metacognition
Signal detection theory

ABSTRACT

In this study we investigated the influence of list composition on judgments of learning (JOLs). To this end, we compared JOLs assigned in a multi-cycle procedure to a set of moderately difficult word pairs. Experiment 1 revealed that when difficult new pairs were added to the study list, the mean of JOLs assigned to the moderate pairs increased as compared to the baseline. In Experiment 2, we reversed this pattern by including easy new pairs in the study list. By analyzing metacognitive ROCs (MROCs), we demonstrate that these results were caused by criterion shifts, by which participants adjusted the level of evidence needed to assign particular JOL ratings. Changes in the study list composition led to a recalibration of the JOL scale – i.e. resetting of the criteria – in order to accommodate the addition of new items. We discuss the usefulness of MROCs for detecting criterion shifts in rating tasks.

Introduction

Rating scales are ubiquitous in psychological research. In general, the scales used by psychologists can roughly be divided into two groups (e.g., Biernat, Manis, & Nelson, 1991; Frederick & Mochon, 2012). *Subjective scales* are characterized as having no predetermined meaning: the interpretation of the points on these scales cannot be inferred *a priori*, without taking into account what the ratings actually refer to. For example, on a scale ranging from *very small* to *very large*, the precise meaning of the labels depends on the range of sizes of to-be-rated items. With such scales, there is no contradiction that a *very small* mammal can still be larger than a *very large* insect. *Objective scales*, on the other hand, have predefined, objective referents. The interpretation of, say, weight in grams should always be the same, independent of whether the animal being weighed is an insect or a mammal.

In memory and metamemory research, researchers commonly use measures such as retrospective confidence (RC) judgments, and prospective measures such feeling-of-knowing (FOK) judgments or judgments of learning (JOLs), amongst others, to investigate internal assessments of participants' own knowledge. Often the scales metacognitive theorists use are subjective, such as a *1*-to-*6* scale of RC.[1] Metacognitive studies employing subjective scales are often concerned with *resolution* – that is, the extent to which the assigned scale values discriminate between correct versus incorrect responses on some criterial test (e.g., correctly recalled vs. not correctly recalled on a recall test following a JOL judgment; correctly recognized vs. not correctly recognized on a recognition test following an FOK judgment, etc.). For resolution, the absolute magnitude of judgments

☆ The authors would like to thank Maciej Hanczakowski and Greg Neil for their helpful comments concerning this research.

* Corresponding author at: Division of Psychology, Nottingham Trent University, Burton Street, Nottingham NG1 4BU, UK.
E-mail address: katarzyna.zawadzka@ntu.ac.uk (K. Zawadzka).

[1] Throughout the paper, scale labels are italicized.

http://dx.doi.org/10.1016/j.jml.2016.04.005
0749-596X/

is irrelevant, as long the ratings distinguish correctly between these two types of responses. So, for example, if a person assigned FOK ratings of *6* to all subsequently recognized items, the same perfect resolution would be obtained as long as they assigned any ratings lower than *6*, be it *5* or *1*, to all subsequently unrecognized items. Popular measures of resolution, such as gamma correlations or signal detection measures of d', d_a, or area under the Receiver Operating Characteristic (ROC) curve can be calculated from an ordinal scale, and a subjective *1*-to-*6* scale satisfies this requirement.

The same metacognitive ratings can also be elicited on objective scales, such as *0–100%* scales of subjective probability. In order for this scale to be interpreted as objective, the scale values must have some pre-set referents. It is assumed that they refer to the likelihood of some outcome in the long run (a frequentist approach to probability). In the case of JOLs, a rating of *40%* would mean, then, that a person predicts recalling at a future test 40% of all items assigned this rating.

Objective metacognitive scales have one notable advantage over their subjective counterparts: they allow for an additional measure of metacognitive accuracy to be calculated which reflects the correspondence between ratings and objective performance: *calibration*. Calibration can be assessed at separate levels on the rating scale (e.g., percentage correct is calculated separately for all items assigned a rating of *0–9*, *10–19*, ..., *90–99*, *100%* and then ratings and percentage correct are compared at each level), or for the whole test. In both cases, perfect calibration (or realism) requires that the means corresponding to objective performance are equal to mean ratings assigned to the items. On the other hand, a rating mean that is lower than the performance mean is interpreted as underconfidence, whereas the reverse pattern is interpreted as overconfidence. Therefore, it is assumed that by having participants use the objective *0–100%* JOL scale, researchers can gain insight into how good they are at estimating, in objective terms, their overall level of knowledge. Calibration scores have been used by experimenters to draw conclusions about potential similarities or differences in monitoring abilities in developmental research (e.g., Connor, Dunlosky, & Hertzog, 1997; Lipko, Dunlosky, Lipowski, & Merriman, 2012; Rast & Zimprich, 2009), eyewitness research (e.g., Allwood, Ask, & Granhag, 2005; Sauer, Brewer, Zweck, & Weber, 2010) and educational research (e.g., Butler, Karpicke, & Roediger, 2008; Dunlosky & Rawson, 2012), among many other areas of psychology.

However, some concerns regarding the interpretation of the *0–100%* JOL scale have been formulated in the JOL literature. Recently, Hanczakowski, Zawadzka, Pasek, and Higham (2013; see also Higham, Zawadzka, & Hanczakowski, 2016; Zawadzka & Higham, 2015) cast doubt on the likelihood interpretation of percentage JOLs. Their research concerned the underconfidence-with-practice (UWP) effect (see, e.g., Finn & Metcalfe, 2007, 2008; Koriat, Sheffer, & Ma'ayan, 2002), an impairment of calibration present when the same materials are studied and tested more than once. In a typical UWP experiment, participants first study a list of (typically unrelated) cue–target pairs such as *digit-hunger*. During study, they assign JOLs to each item to indicate how likely it is that they will later remember the target of the pair if provided with the cue on an immediate cued-recall test following study. Following the list, a recall test is administered and performance on this test is compared to JOLs assigned during study. On this first test, participants are typically well calibrated or there is slight overconfidence. Following the first test, the entire procedure is repeated at least once so that the whole experiment consists of two or more identical study-test *cycles*. However, unlike the results from the first cycle, from the second cycle onward, participants are typically underconfident; that is, their JOLs underestimate their actual recall.

Hanczakowski, Zawadzka, et al. (2013) noted that the UWP effect was independent of the instructions given to participants regarding the interpretation of JOLs. In most studies participants were cued at study with a prompt asking them to rate the likelihood of recalling the target at test, such as "With what probability will you remember the target word in about five minutes from now if you see the cue word?" (Rast & Zimprich, 2009). Instructions like these should, at least in theory, convey to participants that the JOL task is in fact a probability rating task, and so the JOL scale is an objective one, with JOL values indicating assessed probability of recall. However, some researchers have used JOL prompts that did not mention the constructs of probability or likelihood at all, and asked instead about confidence (e.g., Scheck & Nelson, 2005; Serra & Dunlosky, 2005). Nevertheless, despite the fact that the likelihood and confidence prompts are profoundly different on a theoretical level, there was no difference in the accuracy (as assessed by calibration) of likelihood- and confidence-prompted JOLs.[2] This led Hanczakowski, Zawadzka et al. to question whether participants in the percentage JOL task were really aiming to maximize calibration. If they were not, this would be consistent with findings from the judgment and decision making literature suggesting that participants do not aim at assessing calibration even if they are provided with direct instructions to do so and examples of what calibration entails (Keren & Teigen, 2001; Lichtenstein & Fischhoff, 1981).

For this reason, Hanczakowski, Zawadzka, et al. (2013) decided to assess the generalizability of the UWP effect to different rating types, such as binary *yes*/*no* JOLs and binary betting decisions.[3] They argued that if the UWP effect was found with ratings other than *0–100%* JOLs, it would be consistent with the claim that this effect reflects inaccurate assessments of the likelihood of future recall. However, what Hanczakowski, Zawadzka et al. found is that, in contrast to the underconfidence observed with the percentage-JOL scale, the proportion of "yes" responses on later cycles with the binary tasks did not differ from the proportion of correctly recalled items, revealing good calibra-

[2] Luna, Higham, and Martín-Luengo (2011) observed similar correspondence between likelihood ratings and RC ratings in a retrospective task.

[3] With binary tasks, realism would be evident if the percentage of "yes" responses (i.e., binary JOL: "yes, I will remember the item later"; binary betting: "yes, I am willing to bet that I will recall the item later") equaled the percentage of items actually recalled.