Judgment and Decision Making, Vol. 1, No. 1, July 2006, pp. 86–90

The availability heuristic in the classroom: How soliciting more criticism can boost your course ratings

Craig R. Fox*
UCLA Anderson School and Department of Psychology

Abstract

This paper extends previous research showing that experienced difficulty of recall can influence evaluative judgments (e.g., Winkielman & Schwarz, 2001) to a field study of university students rating a course. Students completed a mid-course evaluation form in which they were asked to list either 2 ways in which the course could be improved (a relatively easy task) or 10 ways in which the course could be improved (a relatively difficult task). Respondents who had been asked for 10 critical comments subsequently rated the course more favorably than respondents who had been asked for 2 critical comments. An internal analysis suggests that the number of critiques solicited provides a frame against which accessibility of instances is evaluated. The paper concludes with a discussion of implications of the present results and possible directions for future research.

Keywords: availability, course evaluations, accessibility, easy of retrieval

1 Introduction

According to Tversky and Kahneman's (1973) availability heuristic, people sometimes judge the frequency of events in the world by the ease with which examples come to mind. This process has generally been demonstrated by asking participants to assess the relative likelihood of two categories in which instances of the first category are more difficult to recall than instances of the second category, despite the fact that instances of the first category are more common in the world. For instance, Kahneman and Tversky (1973) found that most people think the letter *R* more often appears in English words as the *first* letter than the *third* letter, presumably because the first letter provides a better cue for recalling instances of words than does the third letter. In fact, it turns out that *R* appears more often as the third than first letter in English words.

Schwarz *et al.* (1991) observed that the classic studies demonstrating the availability heuristic failed to distinguish an interpretation based on *ease* of retrieval from an alternative interpretation based on *content* of retrieval in which an event is judged more common when a larger number of examples come to mind. To tease apart these accounts, Schwarz *et al.* (1991) asked participants in one study to list either 6 or 12 examples of assertive or unassertive behavior that they have exhibited and then rate themselves on their overall degree of assertiveness. Participants rated themselves as more assertive after they had listed 6 examples of assertive behavior (a relatively easy task) rather than 12 examples (a relatively difficult task); similarly, they rated themselves as less assertive (i.e., more unassertive) after they had listed 6 rather than 12 examples of *unassertive* behavior. Similar patterns of results have been observed in many other studies of frequency-related judgments, including the rate at which a particular letter occurs in various positions of words (Wänke, et al., 1995), the quality of one's own memory (Winkielman, et al., 1998), the frequency of one's own past behaviors (Aarts & Dijksterhuis, 1999), one's susceptibility to heart disease (Rothman & Schwarz, 1998) and one's susceptibility to sexual assault (Grayson & Schwarz, 1999). For a review of this literature see Schwarz (1998; 2004). Thus, an abundance of data supports the original interpretation of the availability heuristic: categories are judged to be more common when instances more *easily* come to mind, even when a smaller absolute *number* of instances are generated.

This program has been extended from frequency-based judgments to evaluative judgments of such targets as public transportation (Wänke, et al., 1996), luxury automobiles (Wänke, et al., 1997), and one's own childhood (Winkielman & Schwarz, 2001). For instance, Winkielman and Schwarz (2001) asked participants to recall either 4 childhood events (an easy task) or 12 childhood

*I thank Jim Bettman, Rick Larrick, Patty Linville and Yael Zemack for helpful comments and suggestions on an earlier draft of this paper. Address correspondence to: Craig R. Fox, UCLA Anderson School, 110 Westwood Plaza #D511, Los Angeles, CA 90095–1481, craig.fox@anderson.ucla.edu

events (a difficult task). Some participants were then led to believe that memories from pleasant periods tend to fade, while others were led to believe that memories from unpleasant periods tend to fade. When later asked to *evaluate* their childhood, participants believed that pleasant memories fade rated their childhood more favorably when they completed the difficult task (12 events) than the easy task (4 events); participants who believed that unpleasant memories fade rated their childhood more favorably when they completed the easy rather than difficult task.

Previous studies of the availability heuristic using the paradigm of Schwarz *et al.* (1991) have turned up impressive and robust results. However, these demonstrations have been restricted primarily to laboratory surveys in which task of recalling examples then making an overall assessment may seem somewhat artificial to participants and the responses of little consequence. More important, most participants in previous studies presumably had little prior experience with the particular Likert scale that served as the dependent measure (e.g., most had never before rated their childhood or public transportation on a 7-point scale). Hence, ratings of respondents may be especially susceptible to superficial cues—such as the accessibility of instances—when mapping their beliefs and attitudes onto an unfamiliar response scale.

The present investigation overcomes these limitations through a "field study" of students evaluating a course. First, evaluations are a normal facet of most university courses in which students are commonly asked to list specific suggestions and also provide a global assessment. Moreover, course evaluations are consequential, as they can influence future course offerings and course staffing, promotion and tenure decisions, and provide information to future prospective students of the target course. Second, students at universities quickly become familiar with standard course evaluation scales and how ratings are distributed across classes, often relying on these scores in choosing among elective courses.

The study of course evaluations is also interesting in its own right. A number of recent papers have questioned the validity of these ratings, and a lively debate appeared some years ago in the *American Psychologist* (1997; pp.1182-1225; 1998, pp.1223-1231). Thus far, questions of discriminant validity have mainly focused on the correlation between teaching ratings and apparently irrelevant factors such as the students' expected grades or the course workload. To date there have been few published investigations of the relationship between the design of course feedback forms and summary course evaluations. The present study attempts to answer the following provocative question: can one paradoxically obtain *higher* course ratings by soliciting *a greater number* of critical comments from students?

2 Method

Participants were 64 business students enrolled in two sections of a course on negotiation at Duke University. Three weeks into a six-week term, students were asked to complete a one-page mid-course evaluation form, as they do for most classes in the business school at Duke. Students in both class sections were randomly assigned to one of two course evaluation forms that differed by a single item. The first ten items on both forms were neutrally valenced short-answer and multiple choice questions (e.g., "How do you view the pacing of the class," with response options ranging from "much too slow" to "much too fast") and neutrally valenced open-ended questions (e.g., "What do you think of the lectures and class discussion so far?"). For item #11, half the students (n = 32) were asked to "List 2 ways in which you feel the course could be improved" whereas the remaining students (n = 32) were asked to list 10 ways in which they felt the course could be improved. Directly below this question, the relevant quantity of numbered spaces (2 or 10) were provided. For item #12, all participants were asked to list their 2 favorite aspects of the course. Finally, all participants were asked, "Overall, how would you rate the course so far on a 1-7 scale."[1] This scale is used for all final course evaluations, with 1 denoting the lowest possible score and 7 the highest possible score. The large majority of students enrolled in the class (84%) were in the midst of their second year of the MBA program (their seventh six-week term), and therefore had an abundance of prior experience using this scale to rate classes.

3 Results

Responses did not differ significantly between the two class sections, so the data were combined. Results supported the major theoretical prediction: course evaluations were higher among students who were asked to list 10 ways in which the course could be improved (M = 5.52, SD = 0.88) than among those were asked to list 2 ways in which the course could be improved (M = 4.92, SD = 1.12). Median scores were 5.5 and 5.0, respectively. These differences were statistically significant ($t(56)$ = 2.24, p = .03). Based on the distribution of final scores for all courses offered that term in the business school at Duke University, this difference between experimental conditions is approximately $\frac{3}{4}$ of a standard deviation.[2]

[1] Six respondents provided course critiques but no summary course rating; hence their responses were dropped from the sample for analyses in which the latter variable entered.

[2] The average final course rating among all classes taught in the business school that term was 5.34, SD = 0.80. The course reported in this study received a final rating of 5.82.

https://doi.org/10.1017/S1930297500000371 Published online by Cambridge University Press

Table 1: Results of linear regressions predicting course evaluation scores.

	Model Number						
	1	2	3	4	5	6	7
Critiques solicited	0.0749* (0.0334)			0.0855* (0.0334)	−0.00172 (0.0446)		−0.00558 (0.0632)
Critiques produced		−0.118 (0.0957)		−0.162† (0.0928)		0.00344 (0.0971)	0.012 (0.138)
Ratio of produced to solicited			−1.070** (0.312)		−1.081* (0.439)	−1.075** (0.343)	−1.12† (0.667)
Adj R^2	.066	.009	.159	.099	.144	.144	.128

† $.05 < p < .10$, * $p < .05$, ** $p < .005$

Note: Rows correspond to independent variables. Regression models are numbered by column. Cell entries list regression coefficients, with standard deviations in parentheses. Adjusted R^2 values for each regression are listed below the relevant column.

Not surprisingly, participants produced a greater number of suggestions in the 10-slot condition (M = 2.1; SD = 1.84) than in the 2-slot condition (M = 1.6, SD = 0.72); however, this difference was small and not statistically significant ($t(56) = 1.52$, $p = .14$). In fact, only 31% of participants in the 10-slot condition provided more than two suggestions for ways in which the course could be improved. One might therefore surmise that the observed effects on course ratings were more often driven by artificially *extending* the natural retrieval process of students in the 10-slot condition (thereby improving subsequent course ratings) than by artificially *truncating* the natural retrieval process of students in the 2-slot condition (thereby depressing subsequent course ratings).

In previous studies using similar methods, the vast majority of participants did manage to produce the number of examples that were solicited by the experimenter so that the number of examples solicited and the number provided were perfectly (or nearly perfectly) correlated. In the present study, the variability in the number of critiques provided by respondents in both conditions offers a unique opportunity to explore the independent effects of the number provided and the number solicited.[3] When both variables are entered into an ordinary least-square regression, the number *solicited* is positively related to course ratings ($b = .09$, $t(55) = 2.56$, $p = .01$), and the number *produced* is negatively related to course ratings ($b = -.16$, $t(55) = -1.74$, $p = .09$; see model 4, Table 1).

One might infer on the basis of this result that the number of critiques solicited provides a frame against which the number recalled is evaluated. If so, one would expect course ratings to be predicted quite well by the *ratio* of critical comment produced to critical comments solicited. Indeed, there is a strong negative correlation between this ratio and course ratings ($r = -.42$; $t(56) = -3.43$, $p < .001$; see also model 3, Table 1). Moreover, when we enter both the number of critiques solicited and the ratio of critiques produced to solicited into a regression (model 5, Table 1), the independent effect of the number solicited is wiped out ($b = 0.00$, $t(55) = 0.04$, $p = .97$), but the association between course ratings and the ratio of criticisms produced to solicited remains strong and significant ($b = -1.08$, $t(55) = -2.47$, $p = .02$).[4] Similarly, when we enter both the number of critical comments produced and the ratio of critiques produced to solicited into a regression (model 6, see Table 1), the independent effect of the number of critiques produced is wiped out ($b = 0.00$, $t(55) = -0.04$, $p = .97$), but the independent association between course ratings the ratio of critiques produced to solicited remains strong and significant ($b = -1.07$, $t(55) = -3.13$, $p = .003$).[5] Taken together, these results suggest

[3]Participants in the studies of Grayson and Schwarz (1999) also failed to report all the examples solicited, however these investigators report no such internal analysis.

[4]Multicollinearity does not seem to be a problem. The correlation between number of critiques produced and the ratio of critiques produced to solicited is 0.38, and the Variance Inflation Factor (VIF) in the multiple regression is 1.2.

[5]Again, multicollinearity does not seem to be a problem. The correlation between number of critiques solicited and the ratio of critiques

https://doi.org/10.1017/S1930297500000371 Published online by Cambridge University Press

that the *ratio* of criticisms produced to solicited mediates the relationships between both of these variables and course ratings (cf. Baron & Kenny, 1986).

4 Discussion

The present study provides strong evidence that course evaluations can be improved, paradoxically, by soliciting a larger number of critical comments from students. Previous laboratory studies of the availability heuristic have found that a target category is sometimes judged less common after a greater number of examples are solicited. The present study extends this program to evaluative judgments in a naturalistic context in which participants have ample prior experience mapping their attitudes to the relevant response scale.

Most previous studies determined through pilot testing that it would be difficult for participants to provide the total number of examples solicited. However, most studies were either designed so that *all* participants would be able to eventually produce the total number of examples solicited (e.g., Schwarz, *et al.,* 1991) or did not ask participants to explicitly produce any examples but merely *ponder* the task of producing the number of examples solicited (Wänke, Bohner & Jukowitsch, 1997). The present study provides a more direct measure of how difficult each respondent found the task: the number of critiques that the respondent produced.

The finding that the *ratio* of critiques produced to solicited mediates the relationship between the number solicited and course ratings suggests that the number solicited may be adopted as a norm against which the retrieval of critiques is evaluated. That is, the number solicited may be accepted as a "reasonable" or "expected" quantity of criticism, consistent with the rules of conversational implicature (Grice, 1975). Indeed, Winkielman *et al.* (1998) observed the usual pattern of results when participants were told that most people find the task of recalling a large number of examples *easy* (which implies that the number solicited is an appropriate norm), but the reverse pattern when participants were told that most people find the task *difficult* (which implies that the number solicited is not an appropriate norm). Future research might test the "norm" hypothesis more directly by examining whether the association between frequency ratings and number of examples solicited is moderated by the perceived normative diagnosticity of the number of examples solicited. For instance, if participants learn that the number of examples solicited was determined by the roll of dice, one might expect the effect size to diminish; if participants are explicitly told that the task is designed so that an average person can complete it with modest effort, one might expect the effect size to be augmented.

The present investigation demonstrates that a minor variation in the format of course evaluation forms—in this case, changing a single word ("two" to "ten") and changing the number of spaces provided for responses—can have a pronounced effect on global course evaluations that are made on a familiar rating scale. One might expect that other superficial manipulations of format, such as the order in which the global evaluation versus constituent judgments are solicited, may also affect measured course ratings (see, e.g., Sudman, et al., 1992). On a practical level, the present results underscore the importance of standardization of course evaluation forms when summary scores are compared across classes or departments. Additionally, the lability of course evaluations reminds us of the limitation of summary evaluation scores as a measure of teaching performance.

produced to solicited is –0.72, and the VIF is 1.9.

References

Aarts, H. & Dijksterhuis, A. (1999). How often did I do it? Experienced ease of retrieval and frequency estimates of past behavior. *Acta Psychologica, 103,* 77–89.

Baron, R. M. & Kenny, D. A. (1986). The Moderator-mediator variable distinction in social psychology research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51,* 1173–1182.

Grayson, C. E., & Schwarz, N. (1999). Beliefs influence information processing strategies: Declarative and experiential information on risk assessment. *Social Cognition, 17,* 1–18.

Grice, H. P. (1975). Logic and Conversation. In P. Cole, &J. L. Morgan, (Eds.), *Speech Acts*, pp. 41–58. London: Academic Press.

Kahneman, D., & Tversky, A. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131.

Menon, G., Raghubir, P., & Schwarz, N. (1995). Behavioral frequency judgments: An accessibility-diagnosticity framework. *Journal of Consumer Research, 22,* 212–228.

Rothman, A. J., & Schwarz, N. (1998). Constructing perceptions of vulnerability: Personal relevance and the use of experiential information in health judgments. *Personality and Social Psychology Bulletin, 24,* 1053–1064.

Schwarz, N. (1998). Accessible content and accessibility experiences: The interplay of declarative and experiential information in judgment. *Personality and Social Psychology Review, 2,* 87–99.

https://doi.org/10.1017/S1930297500000371 Published online by Cambridge University Press

Schwarz, N. (2004). Metacognitive experiences in consumer judgment and decision making. *Journal of Consumer Psychology, 14,* 332–348.

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., & Simmons, A. (1991). Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social Psychology, 61,* 195–202.

Sudman, S., Bradburn, N. M. & Schwarz, N. (1996). *Thinking About Answers: The application of Cognitive Processes to Survey Methodology.* San Francisco: Jossey-Bass.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5,* 207–232.

Wänke, M., Bless, H. & Biller, B. (1996). Subjective experience versus content of information in the construction of attitude judgments. *Personality and Social Psychology Bulletin, 22,* 1105–1113.

Wänke, M., Bohner, G, & Jukowitsch, A. (1997). There are many reasons to drive a BMW: Does imagined ease of argument generation influence attitudes? *Journal of Consumer Research, 24*, 170–177.

Wänke, M., Schwarz, N. & Bless, H. (1995). The availability heuristic revisited: Experienced ease of retrieval in mundane frequency estimates. *Acta Psychologica, 89*, 83–90.

Winkielman, P., & Schwarz, N., & Belli, R. F. (1998). The Role of ease of retrieval and attribution in memory judgments: Judging your memory as worse despite recalling more events. *Psychological Science, 9*, 124–126.

Winkielman, P., & Schwarz, N. (2001). How pleasant was your childhood? Beliefs about memory shape inferences from experienced difficulty of recall. *Psychological Science, 12*,176–179.