12-18-2010

The Heart and Mind at Work: The Effects of Implicit and Explicit Reasoning on Performance Appraisal

Scott Ryan
University of Connecticut - Storrs, scott.ryan@uconn.edu

The Heart and Mind at Work: The Effects of Implicit and Explicit Reasoning on Performance Appraisal

Scott Ryan

M.S., Brown University, 1999

A Thesis

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Arts

at the

University of Connecticut

2010

APPROVAL PAGE

Master of Arts Thesis

The Heart and Mind at Work: The Effects of Implicit and Explicit Reasoning on Performance Appraisal

Presented by

Scott Ryan

Major Advisor

_____

R. James Holzworth

Associate Advisor

_____

Janet L. Barnes-Farrell

Associate Advisor

_____

Lucy Gilson

University of Connecticut

2010

Acknowledgments

I would to thank Jim Holzworth for all of his help and patience over the past two years. It is greatly appreciated. Janet Barnes-Farrell has been extremely helpful at every stage of my time here at the University of Connecticut. I appreciate Lucy Gilson agreeing to be on my committee. I would also like to that my research assistant, Scott Emch for helping with data collection, and Megan Dove-Steinkamp for helpful suggestions throughout the process.

Table of Contents

The Heart and Mind at Work: The Effects of Implicit and Explicit Reasoning on Performance Appraisal

Since Landy and Farr's (1980) watershed review and evaluation of the performance appraisal literature, research on performance appraisal has been redirected toward investigation of the cognitive processes involved (Ilgen, Barnes-Farrell, & McKellin, 1993). Barnes-Farrell (2001) notes that there are a number of important cognitive processes involved in performance appraisal. Because individuals cannot explicitly process every piece of information about a subordinate, they must employ some cognitive simplification strategies. These simplification strategies are associated with implicit reasoning (Hogarth, 2001) and, according to Payne and Gawronski (2010), evidence is emerging that a great deal of social information processing is implicit. Barnes-Farrell (2001) discusses how implicit reasoning can increase accuracy in performance appraisal. Specifically, appraisers often face information overload; when facing information overload, implicit processes can be more accurate than explicit processes (Hogarth, 2005). McMackin and Slovic (2000) and Wilson and Schooler (1991) explored different task conditions under which implicit and explicit reasoning are superior in different cognitive tasks. The purpose of the present study is to further investigate implicit and explicit reasoning in the context of performance appraisal, under conditions of low and high cognitive load.

The current study investigates the merits of explicit and implicit reasoning processes in performance appraisal judgments. Explicit processes are explicit in the sense that one is aware of the details of the process (MacDonald, 2008; Hogarth,

2001). Implicit processes are implicit in the sense that one is *not* aware of the details of the process. Explicit processes are often associated with controlled and conscious processes, whereas implicit processes are associated with automatic and unconscious processes (Schneider & Shiffrin, 1977).

Explicit processes are associated with analysis whereas implicit processes are associated with intuition (Hammond, 1996, Hogarth, 2001). Hammond's (1996) cognitive continuum theory explains the relationship between analysis and intuition. Analysis is described as being slow, high in cognitive control, and high in conscious awareness, whereas intuition is described as being fast, low in cognitive control, and low in conscious awareness. As the name implies, cognitive continuum theory does not consider a process to be completely analytical or completely intuitive. Analysis and intuition are opposite ends of a continuum. A process can be partly analytical and partly intuitive. The current study examines explicit and implicit processes in performance appraisal judgments. As suggested by cognitive continuum theory, a performance appraisal judgment may be partly explicit and partly implicit. Individuals making performance appraisal judgments may use a combination of explicit and implicit processes. Supervisors who are making performance appraisal judgments may explicitly weigh specific aspects of an employee's performance, or simply rely on their intuitive impression of how well the employee is performing. The current study compares the accuracy of these two approaches.

The accuracy of implicit cognitive processes

A great deal of research has investigated the merits of quick, intuitive, implicit judgments (e.g., Hammond, Hamm, Grassia, & Pearson, 1987; McMackin & Slovic,

2000; Hogarth, 2001; Sloman, 1996). One demonstration of the accuracy of implicit cognitive processes was performed by Bechara, Damasio, Tranel, and Damasio (1997). Participants played a game in which they chose cards from different decks. Each card could result in a positive or negative payout. The goal was to find the deck that had the most positive payouts. Participants began showing implicit awareness *before* they were explicitly aware that the decks were positive. The evidence for this implicit awareness was that participants displayed an increased galvanic skin response before choosing a card from the deck with negative payouts. The skin conductance response indicated an increase in sweat, which was assumed to be associated with negative affect. This negative affect response was interpreted as evidence that participants "knew" that the decks were negative. However, when asked whether they knew which deck was more positive, they responded that they did not. These results indicated that the participants behaved as if they knew which decks were advantageous even though this "knowledge" was not explicit. This is one example in which an implicit cognitive process is more accurate than an explicit cognitive process, assuming that accuracy is defined as the speed with which one can uncover a pattern.

Under certain conditions, explicitly considering specific reasons for a judgment can lead to inaccuracy. Wilson and Schooler (1991) performed a manipulation in which individuals were asked to write out reasons for their judgments *before* making the judgments. *In the current paper this is referred to as the justification manipulation.* This manipulation leads individuals to take a process that is implicit and make the process more explicit. Participants made judgments of the

quality of strawberry jams. Some participants wrote down justifications before making the ratings while others did not. Judgment accuracy was defined as matching the ratings given by experts published in a popular consumer magazine. Those who wrote down their justifications did not match the ratings of experts as accurately as those who did not write down their justifications. Accuracy of judgment was reduced by making reasons for the judgment explicit.

McMackin and Slovic (2000) performed a study using the justification manipulation, in which one is asked to provide reasons for a judgment before making the judgment. They had a group of students rate specific print advertisements on likeability. Other participants were then asked to estimate how each advertisement scored in the survey of students. Accuracy was defined as the extent to which participants' ratings matched ratings from other students at their university. In one condition the participants were told to write down the factors that were likely to influence students' ratings of the ads before they provided judgments. In a control condition participants were simply told to provide ratings. Results indicated that participants who wrote down their reasons were less accurate than participants who did not write down their reasons. This is further evidence that making a process explicit can decrease accuracy.

Why would writing down reasons for a judgment make the judgment less accurate? Wilson and Schooler (1991) suggest that participants in the control condition of their study evaluated the stimuli "fairly optimally" (p. 191). If an individuals' initial implicit reaction is accurate, asking someone to make that reaction explicit can only make the judgment less (or equally) accurate. Accuracy is so high

that it cannot increase, it can only decrease or remain the same. Wilson and Schooler (1991) suggest that justification can change an individuals' natural reaction and "To the extent that their initial reaction was adaptive and functional, this change might be in a less optimal direction" (p. 182). Wilson and Schooler (1991) also suggest that individuals often do not understand the true reasons for their attitudes. When they write out their reasons, these reasons may simply represent what is accessible at the time, rather than reflecting true reasons for their belief. When considering their reasons, these reasons may conflict with their implicit appraisal of an object, and this causes them to make a different rating than if they had not written out their reasons.

The accuracy of explicit cognitive processes

It may seem unlikely that focusing on reasons for a judgment can decrease judgment accuracy. Both research and common sense tell us that careful, explicit judgments have great advantages over implicit judgments (Janis & Mann, 1977). However, the claim being made in the current study is *not* that implicit judgments are always superior to explicit judgments. Implicit judgments will be superior to explicit judgments in specific circumstances (Hogarth, 2005). Perhaps the most important condition is whether a process can be specified by rules (Hammond et al, 1987; Hogarth, 2001). These rules could be logical rules or mathematical formulas. This may be why reasoning about physics is a common example of when intuition leads us astray. Basic motion of objects is well specified by precise rules and mathematical formulas, so using intuition rather than analytical formulas is likely to lead to error. Another condition under which explicit judgments are more accurate than implicit judgment is when there is a limited amount of information (Hogarth, 2005). Hogarth

(2005) states that under low analytical complexity, explicit reasoning is more accurate than implicit reasoning.

The McMackin and Slovic study (2000) described earlier found that when rating the likeability of print advertisements, implicit judgments were more accurate than explicit judgments. However, the authors did not claim that thinking about a judgment *always* leads to errors. The advertising task was chosen because it was thought to be a task that would benefit from implicit reasoning, based on criteria specified in cognitive continuum theory (Hammond, 1996). In contrast, certain judgments may be made more accurately if explicit thought is put into them. McMackin and Slovic (2000) tested the hypothesis that explicit judgments are sometimes more accurate than implicit judgments. They asked participants to make judgments about numerical quantities, such as the area of the United States. Like the advertising study, participants in this sample were split into a control group and a justification group. In the justification group participants wrote down their reasons for judgments of numerical quantities before making those judgments. Participants who gave reasons for their judgments were *more* accurate than people who did not give reasons. These results indicate that explicit reasoning can outperform implicit reasoning when performing tasks that favor explicit reasoning (Hammond et. al, 1987; Hogarth, 2001). Under some conditions, providing justification before a judgment can have a favorable effect on the quality of the judgment.

Performance appraisal

A number of performance appraisal studies have investigated processes that involve implicit and explicit reasoning, such as the effect of keeping a diary on

performance appraisal (Bernardin & Walter, 1987; DeNisi, Robbins, & Cafferty, 1989; Varma, DeNisi & Peters, 1996). Writing things in a diary is similar to writing out reasons for a judgment; it makes cognitive processes explicit. DeNisi, Robbins and Cafferty (1989) performed a study in which participants either did or did not keep a diary. Participants watched videotapes of carpenters performing their work, and then rated the quality of the work using a Likert scale. The dependent variable in the study was accuracy of the participant's overall ratings of performance. Accuracy was defined as how well the participants' ratings matched the level of performance displayed in the video segments. The performances of the carpenters were deliberately designed to display three levels of performance: low, medium, and high (and testing showed that the videos did display low, medium, and high performance). Results indicated that participants in the diary condition were the more accurate than participants in the no diary condition. Participants in the diary conditions were able to distinguish between high, moderate, and low performance, but participants in the no diary condition were not able to distinguish between these performance levels. The evidence from this study indicated that writing down information before making ratings leads to greater accuracy.

Varma, DeNisi, and Peters (1996) studied the relationship between diary keeping and affect. Participants were supervisors working in a large multi-national electronics firm. Participants either did or did not keep a diary. Affect was conceptualized in a similar way to likeability, with measures such as "I would like to spend more time with this person." Contrary to the researchers' predictions, the

relationship between affect and ratings was *lowest* in the no diary condition. The researchers expected the relationship to be *highest* in the no diary condition.

Performance appraisal systems often require supervisors to both rate performance and give justification(s) for their ratings (Brutus, 2010). For example, a supervisor may give an employee a rating on a 1 to 5 scale indicating the quality of the employee's work, and this numerical rating may also be delivered with a narrative explaining the strengths and weaknesses of the employee's performance. When looking at a rating and written justification, subordinates may assume that the rating is based on the reasons given in writing. They may assume that their supervisor carefully considered all aspects of their performance, and wrote down all of these reasons; then, after carefully considering all aspects of performance, the supervisor made the numerical rating. However, there is another possibility. Rather than the written narrative influencing the rating, it is also possible that the numerical rating influenced the written narrative. Supervisors are likely to have been watching employees throughout the evaluation period. Over that time, it is likely that they formed judgments of their employees' performance. When it comes time to review the employees' performance, they may already know what numerical ratings they will assign. Therefore, they might first write down a numerical rating, and then develop written justification based on their rating. It is possible that supervisors' numerical ratings conform to the written data they have provided, but it is also possible that they selectively present written justifications that conform to the ratings they have assigned. Results from Wilson and Schooler (1991) and McMackin and Slovic (2000) indicate that writing down reasons before (as opposed to after) making a

judgment can affect the accuracy of the judgment. It is important to understand any effect that order (written narrative first vs. numerical rating first) has on the performance appraisal process.

Memory-based and online judgments

Other performance appraisal studies that have investigated the implicit/explicit distinction have involved the relationship between memory and judgment (DeNisi & Peters, 1996; Sanchez & De La Torre, 1996; Woehr & Feldman, 1993). These studies are similar to the explicit justification studies (McMackin & Slovic, 2000; Wilson & Schooler, 1991) because people write down recalled behaviors before making ratings. Many of these studies are based on the distinction between *memory-based* and *online* judgments (Hastie & Park, 1986). Online judgments occur when individuals are asked to form judgments before they observe a person. These judgments are made "online" in the sense that individuals are forming judgments *while* they are making their observations. Memory-based judgments occur when someone is asked to make a judgment only *after* observing a person. These judgments are memory-based because one must rely on memory when making the judgment. For example, someone may listen to a conversation and then later be asked to make an unexpected judgment, such as an individual's age. Because one did not expect to make the judgment, one must essentially recreate the conversation based on memory, and then make the judgment. The judgment will be based on whatever cues the individual can remember. If one were told *before* listening to the conversation that one would be guessing an individual's age, one would make the judgment while listening to the conversation, and would not need to rely on memory. This is deemed

to be an online judgment. Hastie and Park (1986) showed that when individuals made memory-based judgments, there was a significant correlation between the favorability of recalled behaviors and the favorability of judgments, because the judgment was based on the recalled behaviors. When an online judgment was made, the correlation between the favorability of recalled behaviors and the favorability of judgment was not significant.

When executing a performance appraisal study, it is important to consider whether participants will be making online or memory based judgments, because participants may make different judgments depending on whether the judgments are made online or are memory-based (Woehr & Feldman, 1993). Based on the work of Hastie and Park (1986), Woehr and Feldman (1993) investigated the relationship between performance appraisal judgments and memory. Participants rated the effectiveness of an economics professor delivering a lecture based on a videotape of performance. The memory task involved writing out all of the behaviors that participants could recall from watching the videos. A ratio was calculated between the number of positive behaviors recalled and the number of negative behaviors recalled. This allowed for a correlation to be computed between ratings and memory: The positivity of the rating was compared to the positivity of the memories. The order of the ratings task and memory task was manipulated. One group made the rating first and then performed the memory task, and one group performed the memory task and then made the rating. In the "memory task first" condition, participants wrote about behaviors relevant to the judgment they were about to make, and then made the judgment. This is similar to the justification manipulation used in

the McMackin and Slovic (2000) and Wilson and Schooler (1991) studies, in which individuals write out reasons for a judgment before making the judgment. In the "rating task first" condition of Woehr and Feldman (1993), participants first made judgments, and later tried to recall behaviors displayed in the videos. Because the memory task came after the rating, it could not have influenced the rating. This is similar to the control groups from the McMackin and Slovic (2000) and Wilson and Schooler (1991) studies. In the control groups of these studies, participants did not write out justifications before making ratings.

In the *memory-based judgment* condition of Woehr and Feldman (1993), participants were told to learn as much of the lecture material as possible. In the *online judgment* condition, participants were told to focus on the professor's performance. The results from Woehr and Feldman (1993) are displayed in Table 1. Within the memory-based judgment condition, the order of the tasks did not affect the correlation between memory and rating. In both the "rating task first" and "memory task first" conditions, there was a significant positive correlation between memory and the performance appraisal rating. This correlation was positive because individuals who remembered more positive behaviors made more positive ratings. The memory-rating vs. rating-memory order did not affect the correlation between memory and rating because the same basic process was used to make the performance ratings, whether the memory task was performed first or the rating task was performed first. Within the online judgment condition, the order of the tasks did affect the correlation between memory and rating. When the memory task came *before* the rating, there was a correlation between memory and rating. However, if

the memory task came *after* the rating, there was no correlation between memory and rating. The reason that the order was important was because if the memory task came first, it was able to affect the judgment.

The manipulation used in the current study is similar to the memory-rating order manipulation in Woehr and Feldman (1993). In the current study individuals will watch video segments and rate performance of the individuals in the segments. In the explicit condition individuals will write out their reasons for a judgment before making the judgment. This condition is similar to the condition in Woehr and Feldman (1993) in which individuals first tried to recall behaviors and then made judgments. In the implicit condition of the current study, individuals will not write out reasons before making judgments. This condition is similar to the condition in Woehr and Feldman (1993) in which individuals made ratings before recalling behaviors. In the current study, the hypothesis is that there will be a difference between the implicit and explicit conditions. Woehr and Feldman (1993) found a difference between the rating first and memory first conditions in their study, but only when an online judgment was made. Because of the similarity between the conditions in the current study and the conditions in Woehr and Feldman (1993), it is likely that a difference will be found between the implicit and explicit conditions *only* if individuals are making online judgments. Therefore, in the current study, individuals will be led to make online rather than memory based judgments. This will be accomplished by telling individuals, before they begin watching the video segments, that they will be rating the performance of the individuals in the videos.

Confidence

Under certain conditions, when individuals focus on reasons for a judgment before providing the judgment, the judgment decreases in accuracy (McMackin & Slovic, 2000; Wilson & Schooler, 1991). One possible mechanism for this decrease in accuracy involves confidence. Individuals may doubt themselves after thinking about their reasons for making a judgment (Hammond et. al., 1987; Hogarth, 2001, Tordesillas & Chaiken, 1999). This decrease in confidence may lead them to change their judgment, even though the judgment may have been initially correct. This decrease in confidence is especially likely because research (Hammond et. al., 1987; Hogarth, 2001, Tordesillas & Chaiken, 1999) has indicated that individuals often have more confidence in implicit reasoning than explicit reasoning.

Tordesillas & Chaiken (1999) investigated confidence using the justification procedure similar to McMackin and Slovic (2000). The manipulation had individuals rate the importance of multiple factors that went into their ratings. This manipulation was supposed to make reasoning explicit. The results indicated that those who rated multiple factors were less confident than those who did not. This is consistent with the claim made by Hammond et al. (1987) that those who use an explicit method are less confident than those who use an implicit method.

Preferences in performance appraisal ratings

When rating an individuals' performance, raters are often required to provide both ratings and justification for ratings. If the order of these two processes is important, a related question is which order raters *prefer*. A search of the performance appraisal literature did not yield any research addressing whether

individuals prefer providing ratings before justification or justification before ratings. However, studies that address similar questions (Kelley, 2006; Arkes, 2003) suggest that individuals may prefer making ratings first.

Kelley (2006) studied the response rate of surveys using different types of rating scales. Two different surveys were mailed. Each had the same number of items. One group received a survey that had a single holistic Likert rating scale next to each item. The Likert scale measured the relevance of each behavior queried in the survey. A second group received a scale that disaggregated the rating into frequency and importance items. The frequency and importance items were placed next to one another on each page, so that the actual number of items did not change in the two conditions. The results indicated that people who were mailed the holistic measure had higher response rates than people who were mailed the disaggregated measure. This lower response rate may have occurred because there were twice as many ratings to be made. However, it is also possible that individuals simply have an aversion to making disaggregated ratings.

Arkes (2003) described suggestions that he made to the National Science Foundation and the National Institute of Health. One of the suggestions made to both agencies was to disaggregate their ratings of grant proposals. The grants had been rated in a holistic manner, with each grant receiving a single rating of quality. Arkes suggested that they rate each proposal on four separate criteria, and then average the ratings. He noted that research supported the idea that the averaged ratings would be more valid and reliable than the holistic ratings. However, the disaggregation was not accepted by either organization. The officials at the organizations stated that they

preferred to evaluate with their "gut." They preferred to make a holistic, single rating, rather than carefully breaking down their judgment into various parts. The officials also stated that they did not want to adhere to specific criteria. These data indicate that people prefer to make holistic, intuitive evaluations, rather than ratings based on a number of explicit factors.

Hypotheses

The goal of this study is to investigate explicit and implicit performance appraisal judgments. Researchers have suggested that explicit processes are more accurate than implicit processes when a judgment can be decomposed and rules can be easily discovered (see Hogarth, 2005). A judgment that is based on specific criteria may be more formulaic than a judgment of overall performance. Therefore explicit judgments may be more accurate than implicit judgments when rating a specific behavior.

Hypothesis 1:

When rating a specific aspect of performance, participants will be more accurate when making explicit judgments than implicit judgments. When rating overall performance, participants will be more accurate when making implicit judgments than explicit judgments.

Researchers have also suggested that implicit processes, relative to explicit processes, increase in accuracy as information increases (Hogarth, 2005).

Hypothesis 2:

Under low cognitive load, participants will be more accurate when making explicit judgments than implicit judgments. Under high cognitive load, participants will be more accurate when making implicit judgments than explicit judgments.

Tordesillas & Chaiken (1999) and Hammond et. al. (1987) have indicated that individuals have more confidence in their implicit judgments than their explicit judgments.

Hypothesis 3:

Participants in the implicit condition will be more confident in their ratings than participants in the explicit condition.

Based on the Arkes (2003) and Masicampo and Baumeister (2008), individuals should prefer to make simple, holistic judgments over complex, analytical judgments.

Hypothesis 4:

When choosing between making a rating first or providing reasons for the rating first, participants will prefer to provide the rating first.

Method

Overview

Participants watched video segments of a male or female food server working at a restaurant. In the high cognitive load condition they watched 18 video segments, and in the low cognitive load condition they watched 6 segments. Participants were then asked to make ratings of the performance of the food server. In the explicit condition, participants were first asked to type out their reasons for making the rating before making the rating. In the implicit condition, participants were not asked to

type out their reasons. In the overall condition, participants rated the overall performance of the server, and in the specific condition, participants rated the specific behaviors of memory and cooperation. Participants then answered questions to assess how well they remembered behaviors displayed in the video segments.

Participants

Data were collected from 316 undergraduate students enrolled in introductory psychology classes. Participants received course credit for their participation. The sample included 203 women and 113 men. A total of 74 participants had worked as food server and 242 had not.

Design

This experiment was a 2 x 2 x 2 completely crossed between-subjects design. Participants were randomly assigned to one of eight conditions. The factors were justification type (explicit, implicit), cognitive load (low, high) and rating type (overall, specific).

Materials

All data in the study were collected with *Media Lab* version 2006. Media Lab is a software tool designed to present stimuli and measure responses in behavioral science studies. The materials used in this study were a series of video segments. They presented a male or female server working at a restaurant. A sample script is displayed in Appendix A. The video segments were originally developed by Barnes-Farrell (1984). Lewis (2006) recreated these tapes in order to make them appear more contemporary. Half of the videos portrayed a female food server named Karen, and half portrayed a male food server named Mike. There are two versions of each

video, one with Mike and one with Karen. Except for the different food servers, the two versions of each video are the same.

In the high cognitive load condition there were 12 *target* video segments and 6 *distracter* video segments. In the low cognitive load condition there were 4 target video segments and 2 distracter video segments. Participants in the high load condition rated 12 target video segments in the first part of the study, and 6 distracter segments at the end of the study. Participants in the low load condition rated 4 target video segments in the first part of the study, and 2 distracter segments at the end of the study. Table 2 and Table 3 display the behaviors featured in the videos segments: memory, dinner bill activities, maintaining performance levels, and cooperation. The video segments were played one after another with no pauses between segments. For half of the participants the target video segments included the male server and distracter segments included the female server. For the other half of participants target video segments included the female server and distracter segments included the male server. The segments were arranged randomly for each participant.

Expert scores

As suggested by Murphy, Garcia, Kerkar, Martin, and Balzer (1982) and Borman (1978), accuracy of performance appraisal ratings was defined as the difference between participants' ratings and "expert scores." The procedure for obtaining expert scores was based on the method used by Borman (1978). Four (two men and two women) Ph.D. candidates in Industrial and Organizational Psychology each rated 18 videos. All of these subject matter experts had worked as food servers. The experts watched the videos on computer monitors. Each video was rated

individually, with one rating for each video. Because the videos of the male and female server were identical, experts rated only the videos of the female. Experts were informed of potential rating errors, including halo, leniency, and contrast effects. They were each given a piece of paper that explained the criteria for each of the six performance dimensions (see Appendix B).

The intraclass correlation coefficient (ICC) was used as a measure of reliability for the expert ratings. According to LeBreton and Senter (2008), one specific type of ICC, the ICC(A, K), is appropriate when ascertaining the extent to which a mean rating assigned by multiple judges is reliable. In the current study we are interested in the reliability of the mean rating provided by the expert judges. The value of the ICC(18, 4) was .89. This measure is greater than .80, indicating high reliability (LeBreton & Senter, 2008).

Procedure

An overview of the procedure is presented in Table 4. The study was conducted in a computer laboratory. Participants were run in groups ranging from 9 to 18. When participants arrived at the study an experiment information sheet was provided. Participants were informed that participation was voluntary and that they may withdraw at any time. The script in Appendix C was read aloud to the group of participants. Participants were informed that they would be asked to rate the performance of restaurant servers after watching video segments.

After viewing the video segments, participants made ratings of performance on a 7-point Likert scale. Participants rated the server featured in the *target* video segments. Table 5 displays the instructions given as a function of condition (see also

Appendix D). In the explicit condition, participants first typed out their reasons for making ratings before making the performance ratings, whereas in the implicit condition they did not type out ratings. In the overall condition, participants rated overall performance, whereas in the specific condition, they rated the server's cooperation and memory performance. After each performance rating, participants were asked to rate how confident they were about their performance rating on a 7-point Likert scale, ranging from 1 (not at all confident) to 7 (completely confident).

Participants then performed the memory recognition task (Kinicki, Hom, Trost, & Wade, 1995; Lewis, 2006; Lord, 1985). Participants were asked to respond "yes" or "no" to 12 questions asking whether or not certain behaviors occurred. The memory recognition questions are displayed in Tables 2 and 3, and a sample is provided in Appendix E. Half of the correct answers were "yes" and half were "no." The questions in the high load and low load conditions were different because the videos in the two conditions were different.

After performing the memory recognition task, participants performed the final rating (see Appendix F). The purpose of the final performance rating was to determine whether participants preferred to give ratings first or to give reasons for ratings first. Participants were given one of two sets of instructions to counterbalance the order in which the tasks were mentioned. The instructions were: "We want you to give both performance ratings and reasons for the ratings that you give. Please click one of the buttons below" or "We want you to give both reasons for the ratings that you give and performance ratings. Please click one of the buttons below." The screen displayed two buttons that displayed the text "Rating" and "Reasons For

Rating" horizontally. The order was counterbalanced so that in half the cases the "Rating" button was on the left and in half the cases it was on the right.

Participants then filled out the decision making inventory (DMI) (Nygren & White, 2002). The DMI was included as an exploratory measure. The DMI measures the extent to which individuals prefer to use analytical and intuitive reasoning. Finally, participants were asked to provide their gender and whether they ever worked as a food server.

Results

Dependent measure computation

Table 6 displays an overview of computation of the dependent measures. The primary dependent measure was accuracy of performance rating. Accuracy was computed by taking the ratings of participants and subtracting them from the ratings of experts and then taking the absolute value of those scores (as did Borman, 1978, and McMakin & Slovic, 2000). These accuracy scores represent deviation from the expert scores. Because the scores represent deviation from the expert ratings, higher scores represent *less* accuracy.

Recognition accuracy was computed using the $A_{d'}$ measure, a signal detection statistic (Stanislaw & Todorov, 1999). $A_{d'}$ represents participants' ability to distinguish correct and incorrect answers. As suggested by Stanislaw and Todorov (1999, p. 144), hit rate and false alarm rates of 0.0 were replaced with 0.5/n and hit rate and false alarm rates of 1.0 were replaced with (n - .05)/n.

Table 7 displays the mean performance ratings and true (expert) scores for each condition. The performance ratings are participants' ratings of the target videos.

The expert scores displayed in the table are the average of the expert ratings of the target videos. To ensure that expert scores matched the expected level of performance, the average expert score was also computed for low, medium and high performance. The average expert scores were 1.89 for low performance, 4.23 for medium performance, and 7.00 for high performance.

Effects of the justification manipulation on ratings

Table 8 displays mean performance ratings as a function of justification type. To test for the effect of justification for each rating, t-tests were computed separately for the memory, cooperation, and overall ratings. For the memory rating, there was a significant effect for justification, $t(159) = 2.33$, $p < .05$. Participants in the explicit condition gave higher ratings than participants in the implicit condition. For the cooperation rating, there was no significant effect for justification, $t(159) = 2.33$, $p > .05$, and for the overall rating, there was no significant effect for justification, $t(153) = 1.16$, $p > .05$. Only one of the three tests was significant, and the mean effect size for the three mean differences was 0.29. Overall, the justification factor had small effects on performance rating values.

Another way to test the effect of justification is to count the number of words typed. If typing words reflects explicit reasoning, then participants who typed more words may have been using more explicit reasoning than participants who typed fewer words. The correlations between number of words and memory ($r(158) = -.10$), cooperation ($r(158) = .07$), and overall ($r(158) = .09$) ratings were not significant, all ps $> .05$. The number of words that participants wrote did not affect the performance ratings.

Rating accuracy

Table 9 displays scores for rating accuracy and recognition accuracy. Rating accuracy scores represent a combination of overall, memory, and cooperation accuracy ratings (see Table 6 for exact computational details). An ANOVA was performed on rating accuracy. The independent variables for the ANOVA included the three manipulated factors (justification, cognitive load, and rating type) and all two-way interactions between these factors. Participant sex, sex of the server featured in the videos (server sex), and experience as a server were included as control variables. The results of the ANOVA are displayed in Table 10. The ANOVA revealed a main effect for participant sex, $F(1, 306) = 10.71$, $p < .05$. Women ($M = 1.02$, $SD = 0.67$) were significantly more accurate than men ($M = 1.22$, $SD = 0.76$). There was also a main effect for cognitive load, $F(1, 306) = 11.51$, $p < .05$, with high load participants ($M = 0.97$, $SD = 0.70$) being more accurate than low load participants ($M = 1.23$, $SD = 0.69$). This main effect is difficult to interpret because the videos in the two conditions were different, therefore the level of difficulty may have been different in the two conditions. There was a significant interaction between cognitive load and rating type, $F(1, 306) = 14.96$, $p < .05$ (see mean values in Table 11). T-tests revealed that in the low cognitive load condition, participants rating overall performance were more accurate than those rating specific performance, $t(150) = 3.14$, $p < .01$. In the high cognitive load condition, there was no difference between the overall and specific ratings, $t(162) = 1.59$, $p > .05$, although there was a trend toward specific ratings being more accurate than overall ratings.

Hypothesis 1 stated that there would be an interaction between the justification manipulation and the rating type manipulation. Contrary to Hypothesis 1, the interaction between justification and rating type was not significant, $F(1, 306) = 0.74$, $p > .05$. This result indicated that the effect of the justification manipulation on rating accuracy did not differ in the specific and overall conditions. Hypothesis 2 stated that there would be an interaction between justification and cognitive load on rating accuracy. Contrary to Hypothesis 2, the interaction between justification and cognitive load was not significant, $F(1, 306) = 0.56$, $p > .05$. This result indicated that the effect of the justification manipulation on rating accuracy did not differ in the high and low cognitive load conditions. The main effect for justification was also not significant, $F(1, 306) = 1.45$, $p > .05$. The correlation between number of words typed in the explicit condition and rating accuracy was not significant, $r(158) = .08$, $p > .05$.

Recognition accuracy

Table 9 displays scores for recognition accuracy. An ANOVA was performed on recognition accuracy. The independent variables for the ANOVA included the three manipulated factors (justification, cognitive load, and rating type) and all two-way interactions between these factors. Participant sex, sex of the server featured in the videos (server sex), and experience as a server were included as control variables. The results of the ANOVA are displayed in Table 12. There were main effects for both participant sex, $F(1, 306) = 4.84$, $p < .05$, and server sex, $F(1, 306) = 7.70$, $p < .05$ (see Table 13). The main effect for participant sex is further evidence that women were more accurate in the current study. The interaction term between participant sex

and server sex was added to the ANOVA, but was not significant, $F(1, 305) = 2.27$, $p > .05$. There was a main effect for cognitive load, $F(1, 306) = 96.57$, $p < .05$, with low load participants ($M = .81$, $SD = 0.12$) being more accurate than high load participants ($M = .66$, $SD = 0.14$).

Hypothesis 2 stated that there would be an interaction between the justification factor and the cognitive load factor. Contrary to Hypothesis 2, the interaction between justification and cognitive load was not significant, $F(1, 306) = 0.04$, $p > .05$. This result indicated that the effect of the justification manipulation on recognition did not differ in the high and low cognitive load conditions. The main effect for justification was also not significant, $F(1, 306) = 0.04$, $p > .05$. The correlation between number of words typed in the explicit condition and recognition accuracy was not significant, $r(158) = .01$, $p > .05$.

Confidence

The mean confidence ratings are displayed in Table 14. The two confidence values in the specific condition (confidence in memory ratings and confidence in cooperation ratings) were combined using the arithmetic mean. An ANOVA was performed on the averaged confidence values (see Table 15). The independent variables for the ANOVA included the three manipulated factors (justification, cognitive load, and rating type) and all two-way interactions between these factors. Participant sex, sex of the server featured in the videos (server sex), and experience as a server were included as control variables. Hypothesis 3 claimed that participants in the implicit condition would be more confident in their ratings than participants in the explicit condition. The main effect for justification was not significant, $F(1, 306) =$

1.14, $p > .05$. These results do not confirm Hypothesis 3. There was a significant main effect for rating type, $F(1, 306) = 5.63$, $p < .05$. Participants in the specific condition ($M = 6.00$, $SD = 0.78$) were more confident than participants in the overall condition ($M = 5.75$, $SD = 1.02$).

Final rating

Hypothesis 5 claims that when faced with a choice of making a rating first or providing reasons for a rating first, participants will prefer to provide the rating first. A chi square test, $\chi^2(1) = 71.2$, $p < .05$, revealed that more participants chose to provide the rating first (74%) than chose to provide reasons first (26%). Even in the explicit condition, in which participants had been writing their reasons first throughout the experiment, participants chose to provide their rating first 64% of the time, which is significantly greater than those who chose to provide reasons first (36%), $\chi^2(1) = 17.1$, $p < .05$.

When clicking buttons on the screen during the final rating, participants chose the button on the left 72% of the time, which is significantly greater than those who chose the button on the right (28%), $\chi^2(1) = 58.5$, $p < .05$. When the "rating" button was on the left, participants chose that button 97% of the time, which is significantly greater than those who chose the button on the right (3%), $\chi^2(1) = 133.6$, $p < .05$. As a whole, these results are overwhelming evidence that individuals generally tend to choose a button on the left over a button on the right, and that people choose to provide a rating before providing reasons for the rating.

Decision Making Inventory

There were no significant correlations between the DMI and any performance or accuracy ratings (all *p*s > .05).

Discussion

The primary hypotheses in the current study were not supported. Participants who wrote out reasons before making judgments were not more accurate or confident than participants who did not write out reasons. There were no differences between those who did and did not write out reasons on performance rating accuracy, memory recognition accuracy, and confidence in performance ratings. Overall, the effect of writing out reasons was weak.

The effect of the justification manipulation was not only small when making performance ratings, but also when attempting to recall behaviors. The difference on the memory recognition task between those who did and did not type out reasons was nearly zero. One may have suspected that, at the very least, typing out reasons for a judgment would have primed participants' memory for the incidents in the video segments, and recognition accuracy would be higher than those who did not type out ratings. This did not occur, and is evidence of the weak effects of typing out reasons for a judgment. Although prior research has shown that individuals have more confidence in implicit reasoning than explicit reasoning (Hammond et. al., 1987; Hogarth, 2001, Tordesillas & Chaiken, 1999), in the current study there were no differences in confidence between the explicit and implicit conditions. This result is further evidence that writing out reasons has a weak effect on performance appraisal

judgments. It is quite surprising that carefully reflecting on a judgment has such minor effects on judgment, memory recognition, and confidence.

There are several possible explanations for why the effect of the justification manipulation was weak. Whenever a manipulation does not result in a difference between conditions, a natural question to ask is whether the manipulation was strong enough to elicit an effect. The manipulation used in the current study was very strong. It is difficult to imagine a more powerful way of making something explicit than having participants type out their reasons for making a particular rating. Compared to asking participants to think about their reasons and/or to speak their reasons out loud, typing reasons requires them to put their reasoning onto the screen, so they can actually see their reasons clearly. A more likely explanation for the small differences between the explicit and implicit condition involves when judgments were made. Participants were told twice, for emphasis, that they would be watching videos and rating performance of restaurant servers. Based on Hastie and Park (1986), this should have led them to make "online" judgments. These are judgments that are made while participants are watching the video segments. When participants were then asked to type out reasons for their judgments, the judgments had already been made implicitly. In this sense, they are not really writing out the reasons for their judgments *before* the judgments are made, because the judgments had already been made. If this is the case, then typing out reasons may not be helpful or harmful because the judgments have already been made, and it is too late for the additional reasoning to change the judgments.

However, there are reasons to believe that some of the judgments made in the current study were memory based, rather than online, judgments. Memory based judgments are unexpected, which forces individuals to rely on memory to make a judgment (Hastie and Park, 1986). In the specific condition, participants did not know that they would be asked to make ratings specifically about the server's cooperation and memory. When they were asked to make ratings of cooperation and memory, they had to rely on memory to make the judgments. Because there were no significant differences between the explicit and implicit conditions for the specific ratings, it is less likely that the explanation for these small differences is that participants made an online judgment. Woehr and Feldman (1993) found that when individuals made memory based judgments, it did not matter whether they recalled behaviors first or made ratings first, just as in the current study it did not matter whether participants wrote out reasons before making ratings. If participants in the current study made memory based judgments, then when participants were asked to make a performance appraisal rating, they would have consciously considered reasons why they should give a particular rating, even if they were not asked to type out their judgments. If this is the case, then participants in the explicit and implicit conditions were using similar cognitive processes when making ratings. In both conditions, participants explicitly thought about reasons for making ratings. If participants used similar processes in both conditions, then they should have had similar performance ratings, confidence ratings, and memory recognition, as was found in the current study.

Based upon criteria specified in cognitive continuum theory (Hammond, 1996), McMackin and Slovic (2000) assumed that when individuals performed tasks that induce implicit or explicit reasoning, individuals would be more accurate if they utilized a matching mode of cognition. McMackin and Slovic (2000) assumed that rating advertisements would induce implicit reasoning, and they found that individuals were indeed more accurate when using implicit reasoning rather than explicit reasoning. On the other hand, estimating numerical quantities (e.g., area of the U.S. in square miles) should have induced explicit reasoning. Results confirmed their prediction; individuals were more accurate when using explicit reasoning than they were if they used implicit reasoning. Results of the present study do not support those of McMakin and Slovic (2000). Perhaps it is not always true that tasks *inducing* implicit reasoning are performed most *accurately* when using implicit reasoning. For example, although tasks with larger number of attributes (greater than 5) are expected to induce implicit reasoning (Hammond et al, 1987), it is questionable that such tasks will be more accurate when using implicit reasoning than when using explicit reasoning if the compound nature (pattern) of the stimulus attributes does not induce reappraisal. In the current study it is possible that although the performance appraisal task induced implicit reasoning, it was not more accurate when implicit reasoning was used.

Wilson and Schooler (1991) suggest another possible reason for the weak effect of the justification manipulation. They state that individuals are often unaware of why they have certain beliefs. Therefore, when individuals think about the reasons for the beliefs, the reasons they consider may not be the actual reasons. Because

these reasons are different than their actual reasons, beliefs are changed to reflect the written reasons. If this suggestion is correct, then the weakness of the justification manipulation in the current study may have occurred because people *were* aware of the reasons for their beliefs. The videos contained clear examples of negative and positive behaviors, such as showing up late for work, bringing people the wrong orders, remembering everyone's orders, and cooperating with coworkers. Focusing on reasons did not change individuals' ratings because the *true* reasons for their ratings matched their *written* reasons. Wilson and Schooler (1991) also state that if the reasons that individuals consider are the same valence as the actual reasons for the belief, focusing on reasons will not change a belief. In the current study, if individuals formed a positive (or negative) attitude toward the server, and then wrote out positive (or negative) reasons, writing out the reasons would not change the initial attitude/opinion.

Wilson, Kraft, and Dunn (1989) found that when individuals were knowledgeable about an attitude object, there was no attitude change after thinking about the attitude object. Participants in the current study should have been knowledgeable of the job of food server. They are likely to have eaten at restaurants numerous times throughout their lives, and 31% of them had worked as food servers. This familiarity with the subject matter may explain the lack of differences between the explicit and implicit groups. If an unfamiliar job were used instead, maybe there would be differences between the explicit and implicit conditions.

Results from the current study were not consistent with those of earlier studies investigating explicit processes in performance appraisal (DeNisi et al., 1989; Varma

et al., 1996). DeNisi et al. (1989) found that keeping a diary while watching videos of carpenters performing their work led to more accurate appraisal of their performance. Participants in the DeNisi et al. (1989) study were told to write down the tasks performed and how well they were performed while watching the performances. Therefore, they were taking notes *while* watching the videos of performance. In the current study participants wrote down reasons *after* watching the videos of performance. This lack of convergence between the current study and DeNisi et al. (1989) may be due to the fact that in the DeNisi study, individuals could use their diaries to organize information in a meaningful way as they were watching the video segments. In the current study, participants' written accounts were based on memory, which is fallible.

Gender differences

Results indicated that women were more accurate than men on the memory and rating tasks. There are several possible explanations for this effect. It is possible that women are simply more accurate in social judgment. Women are more accurate in decoding non-verbal emotions (Hall, Carter, & Horgan, 2000). Carter and Hall (2008) found the women were more accurate in detecting covariation between group membership and behaviors exhibited by group members, a sign of accurate social perception. Another possibility is that women put more effort into the study. Hyde (2001) reviewed seven studies investigating gender differences in conscientiousness, and indicated that women may be slightly higher in conscientiousness. Four studies indicated women were higher in conscientiousness, two studies indicated that men were higher, and one study found no difference.

Confidence

Prior research has indicated that individuals have more confidence in implicit reasoning than explicit reasoning (Hammond et. al., 1987; Hogarth, 2001, Tordesillas & Chaiken, 1999). In the current study there was no difference in confidence between participants in the explicit and implicit conditions. This suggests that participants in the explicit and implicit conditions may have been using similar cognitive processing.

Results from the current study indicated that participants who made specific ratings were more confident in their ratings than those who made overall ratings. Participants in the specific condition may have been more confident because the precise, detailed directions that were provided in the specific condition may have led participants to believe that they knew what was required of them.

Final rating

The most robust findings in the current study involve the final rating. Participants chose to perform the rating task before providing reasons 74% of the time. Perhaps participants chose to make the rating first because they prefer to make holistic, simple ratings rather than having to first think in detail about the rating (Arkes, 2003; Kelley, 2006; Masicampo & Baumeister, 2008). This preference has a number of possible implications. Individuals may prefer to make holistic ratings in many situations.

Participants chose the button on the left 72% of the time. This may not be surprising given that many things in western culture are ordered from left to right. Santiago, Lupiáñez, Pérez, and Funes, (2007) list things that are ordered from left to

right in Western cultures. These include text, horizontal graphs of time, comic strips, and book pages. Chokron and Agostini (2000) compared the esthetic preferences of French and Israeli participants. French participants read from left to right, but Israeli participants read from right to left. Results indicated that left to right readers preferred pictures facing the right, and right to left readers preferred pictures facing the left. When faced with two buttons on a screen, it appears that the left button is usually the first one that is examined. The current study indicates that the default, or most common, option should be associated with the button on the left.

Conclusion

Based on Wilson and Schooler (1991) and McMackin and Slovic (2000), there should have been differences between the explicit and implicit conditions in the current study. However, these differences did not emerge. The task used in the current study was a performance appraisal rating. This task was different from the tasks used by Wilson and Schooler (1991) and McMackin and Slovic (2000). Further research is needed to determine what characteristics of the current task led to the discrepancy between the current results and the results found in previous research.

Participants either provided justification before making a performance appraisal rating, or simply made the rating. This manipulation was meant to represent real world performance appraisal, in which supervisors provide both a rating and justification for their rating. The results provided evidence that the order of tasks, providing justification first or making ratings first, did not affect performance appraisal ratings strongly. As a practical matter, this may be comforting information,

given that there does not appear to be a common industry standard recommending a particular sequence of ratings and justifications.

References

Arkes, H. R. (2003). Psychology in Washington: The nonuse of psychological research at two federal agencies. *Psychological Science, 14*(1), 1-6. doi:10.1111/1467-9280.01410

Barnes-Farrell, J. L. (1984). The development of a laboratory measure of accuracy in performance appraisal. (Technical Report 84-2) ONR, University of Hawaii.

Barnes-Farrell, J. (2001). Performance appraisal: Person perception processes and challenges. In M. London (Ed.), *How people evaluate others in organizations.* (pp. 135-153). Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science, 275*(5304), 1293-1294. doi:10.1126/science.275.5304.1293

Bernardin, H. J., & Walter, C. S. (1977). Effects of rater training and diary-keeping on psychometric error in ratings. *Journal of Applied Psychology, 62*(1), 64-69. doi:10.1037/0021-9010.62.1.64

Borman, W. C. (1978). Exploring upper limits of reliability and validity in job performance ratings. *Journal of Applied Psychology, 63*(2), 135-144. doi:10.1037/0021-9010.63.2.135

Brutus, S. (2010). Words versus numbers: A theoretical exploration of giving and receiving narrative comments in performance appraisal. *Human Resource Management Review, 20*(2), 144-157. doi:10.1016/j.hrmr.2009.06.003

Carter, J. D., & Hall, J. A. (2008). Individual differences in the accuracy of detecting social covariations: Ecological sensitivity. *Journal of Research in Personality, 42*(2), 439-455. doi:10.1016/j.jrp.2007.07.007

Chokron, S., & De Agostini, M. (2000). Reading habits influence aesthetic preference. *Cognitive Brain Research, 10*(1-2), 45-49. doi:10.1016/S0926-6410(00)00021-5

DeNisi, A. S., & Peters, L. H. (1996). Organization of information in memory and the performance appraisal process: Evidence from the field. *Journal of Applied Psychology, 81*(6), 717-737. doi:10.1037/0021-9010.81.6.717

DeNisi, A. S., Robbins, T., & Cafferty, T. P. (1989). Organization of information used for performance appraisals: Role of diary-keeping. *Journal of Applied Psychology, 74*(1), 124-129. doi:10.1037/0021-9010.74.1.124

Hall, J. A., Carter, J. D., & Horgan, T. G. (2000). Gender differences in nonverbal communication of emotion. In A. H. Fischer (Ed.), *Gender and emotion: Social psychological perspectives.* (pp. 97-117). New York, NY US: Cambridge University Press.

Hammond, K. R. (1996). *Human judgement and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice*. New York, NY US: Oxford University Press.

Hammond, K. R., Hamm, R. M., Grassia, J., & Pearson, T. (1987). Direct comparison of the efficacy of intuitive and analytical cognition in expert judgment. *IEEE Transactions on Systems, Man, & Cybernetics, 17*(5), 753-770.

Hastie, R., & Park, B. (1986). The relationship between memory and judgment depends on whether the judgment task is memory-based or on-line. *Psychological Review, 93*(3), 258-268. doi:10.1037/0033-295X.93.3.258

Hogarth, R. M. (2001). *Educating intuition*. Chicago, IL US: University of Chicago Press.

Hogarth, R. M. (2005). Deciding analytically or trusting your intuition? the advantages and disadvantages of analytic and intuitive thought. In S. Haberstroh (Ed.), *The routines of decision making.* (pp. 67-82). Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

Hyde, J. S. (2001). Gender Differences in Personality and Social Behavior, In: N. J. Smelser & P. B. Baltes (Ed.s), *International Encyclopedia of the Social & Behavioral Sciences*., Oxford, Pergamon (pp. 5989-5994). doi:10.1016/B0-08-043076-7/01784-8.

Ilgen, D. R., Barnes-Farrell, J., & McKellin, D. B. (1993). Performance appraisal process research in the 1980s: What has it contributed to appraisals in use? *Organizational Behavior and Human Decision Processes, 54*(3), 321-368. doi:10.1006/obhd.1993.1015

Janis, I. L., & Mann, L. (1977). *Decision making: A psychological analysis of conflict, choice, and commitment*. New York, NY US: Free Press.

Kelley, J. A. (2006). *The effect of number of rating scales on response rate, response error, and reliability for credentialing job analysis surveys*(ProQuest Information & Learning). *Dissertation Abstracts International Section A: Humanities and Social Sciences, 66*(8-). (2006-99003-127)

Kinicki, A. J., Hom, P. W., Trost, M. R., & Wade, K. J. (1995). Effects of category prototypes on performance-rating accuracy. *Journal of Applied Psychology, 80*(3), 354-370. doi:10.1037/0021-9010.80.3.354

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87*(1), 72-107. doi:10.1037/0033-2909.87.1.72

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*(4), 815-852. doi:10.1177/1094428106296642

Lewis, W.R. (2006). *Effect of Mode of Observation and Level of Performance on*

Accuracy of Performance Appraisal Judgments (Unpublished master's thesis.) University of Connecticut, Storrs, CT.

Lord, R. G. (1985). Accuracy in behavioral measurement: An alternative definition based on raters' cognitive schema and signal detection theory. *Journal of Applied Psychology, 70*(1), 66-71. doi:10.1037/0021-9010.70.1.66

MacDonald, K. B. (2008). Effortful control, explicit processing, and the regulation of human evolved predispositions. *Psychological Review, 115*(4), 1012-1031. doi:10.1037/a0013327

Masicampo, E. J., & Baumeister, R. F. (2008). Toward a physiology of dual-process reasoning and judgment: Lemonade, willpower, and expensive rule-based analysis. *Psychological Science, 19*(3), 255-260. doi:10.1111/j.1467-9280.2008.02077.x

McMackin, J., & Slovic, P. (2000). When does explicit justification impair decision making? *Applied Cognitive Psychology, 14*(6), 527-541. doi:10.1002/1099-0720(200011/12)14:6<527::AID-ACP671>3.0.CO;2-J

Murphy, K. R., Garcia, M., Kerkar, S., Martin, C., & Balzer, W. K. (1982). Relationship between observational accuracy and accuracy in evaluating performance. *Journal of Applied Psychology, 67*(3), 320-325. doi:10.1037/0021-9010.67.3.320

Nygren, T. E. & White, R. J. (2002). Assessing Individual Differences in Decision

Making Styles: Analytical Vs. Intuitive. *Human Factors and Ergonomics Society Annual Meeting Proceedings, Individual Differences in Performance*. (pp. 953-957).

Payne, B. K., & Gawronski, B. (2010). A history of implicit social cognition: Where is it coming from? where is it now? where is it going? In B. K. Payne (Ed.), *Handbook of implicit social cognition: Measurement, theory, and applications.* (pp. 1-15). New York, NY US: Guilford Press.

Sanchez, J. I., & Torre, D. L. (1996). A second look at the relationship between rating and behavioral accuracy in performance appraisal. *Journal of Applied Psychology, 81*(1), 3-10. doi:10.1037/0021-9010.81.1.3

Santiago, J., Lupiáñez, J., Pérez, E., & Funes, M. J. (2007). Time (also) flies from left to right. *Psychonomic Bulletin & Review, 14*(3), 512-516.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. detection, search, and attention. *Psychological Review, 84*(1), 1-66. doi:10.1037/0033-295X.84.1.1

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*(1), 3-22. doi:10.1037/0033-2909.119.1.3

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments & Computers, 31*(1), 137-149.

Tordesillas, R. S., & Chaiken, S. (1999). Thinking too much of too little? the effects of introspection on the decision-making process. *Personality and Social Psychology Bulletin, 25*(5), 623-629. doi:10.1177/0146167299025005007

Varma, A., Denisi, A. S., & Peters, L. H. (1996). Interpersonal affect and performance appraisal: A field study. *Personnel Psychology, 49*(2), 341-360. doi:10.1111/j.1744-6570.1996.tb01803.x

Wilson, T. D., Kraft, D., & Dunn, D. S. (1989). The disruptive effects of explaining attitudes: The moderating effect of knowledge about the attitude object. *Journal of Experimental Social Psychology, 25*(5), 379-400. doi:10.1016/0022-1031(89)90029-2

Wilson, T. D., & Schooler, J. W. (1991). Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology, 60*(2), 181-192. doi:10.1037/0022-3514.60.2.181

Woehr, D. J., & Feldman, J. (1993). Processing objective and question order effects on the causal relation between memory and judgment in performance appraisal: The tip of the iceberg. *Journal of Applied Psychology, 78*(2), 232-241. doi:10.1037/0021-9010.78.2.232

Table 1

Correlations between memory positivity and rating positivity

Judgment type	Memory First	Rating First
Online judgment	0.46*	0.10
Memory-based judgment	0.49*	0.40*

Note. From Woehr and Feldman (1993). A * indicates a significant correlation, $p < .05$. The correlations in the online judgment condition are significantly different from one another.

Table 2

Memory recognition questions in the high cognitive load condition

Performance dimension	Performance level	Memory question	Did the behavior occur?
Memory			
	Low	Karen presented the wrong order to a man and woman and did not apologize for her mistake.	No
	Low	Two men were sitting at a table and Karen gave one man the other man's French fries.	Yes
	Medium	Four men ordered four drinks and Karen asked each person who had which drink and the last drink served was a beer.	Yes
Dinner bill activities			
	Low	Four men were looking at the bill and one of them stated that the bill was wrong because he had not ordered dessert.	Yes
	Medium	Karen brought separate checks to each of three women and asked the first two what they had eaten and then handed the last check to the last woman without asking her what she had eaten.	Yes
	Medium	A coworker noticed that Karen made one mistake on the bill but said that she is always very accurate.	No
Maintaining performance levels			
	Low	Karen called a coworker by her correct name at first but later called her by the wrong name.	Yes
	Medium	A coworker stated that Karen was sweating because she couldn't handle the pressure.	No
	Medium	Two women stated that Karen was frustrated and when she came by she apologized and stated that she was just a little busy that night.	No
Cooperation			
	Medium	Karen offered to help a coworker by telling her to put the full glasses in the middle of the tray.	No
	Medium	A coworker asked Karen to help her and Karen moved some chairs and a table for her.	No
	High	After a male coworker said that there were business men drinking like fish, Karen volunteered to clean his table.	Yes

Table 3

Memory recognition questions in the low cognitive load condition

Performance dimension	Performance level	Memory question	Did the behavior occur?
Memory			
	Low	Karen presented the wrong order to a man and woman and did not apologize for her mistake.	No
	Low	Karen brought the wrong food and the man said that he ordered the angus steak.	No
	Low	Karen brought the wrong food and when she was notified of the mistake she said she would go fix it for them.	Yes
Dinner bill activities			
	Medium	A coworker noticed that Karen made one mistake on the bill but said that she is always very accurate.	No
	Medium	Karen accidentally typed in a discount for carrot cake.	Yes
	Medium	After a coworker told Karen she made an error, Karen said that she did and thanked the coworker.	Yes
Maintaining performance levels			
	Low	Karen called a coworker by her correct name at first but later called her by the wrong name.	Yes
	Low	Karen accidentally grabbed tea instead of coffee.	No
	Low	Karen accidentally called a woman Laurie by the name Linda.	No
Cooperation			
	High	After a male coworker said that there were business men drinking like fish, Karen volunteered to clean his table.	Yes
	High	Karen said that she would clean off his table because she only has a few tables to watch.	No
	High	The man said he had a table of nine and that they were drinking like fish.	Yes

Table 4

Procedure

1. Watch 18 videos (high cognitive load condition) or watch 6 videos (low cognitive load condition)
2. Type reasons for performance rating (explicit condition) or do not type reasons (implicit condition)
3. Rate overall performance (overall condition) or rate cooperation and memory (implicit condition)[1]
4. Rate confidence in the performance rating
5. Answer memory recognition questions
6. Choose order for final rating: Rate performance then type reasons for the rating or type reasons for the rating then rate performance
7. Make final performance rating
8. Fill out decision making Inventory

Note. [1]The memory and cooperation ratings are counterbalanced so that for half of the participants memory is rated first and for half cooperation is rated first.

Table 5
Instructions as a function of condition

	Explicit overall instructions	Explicit specific instructions	Implicit overall instructions	Implicit specific instructions
Screen 1	Overall Instructions + Explicit Instructions	Memory Instructions + Explicit Instructions	Overall Instructions + "Please give a rating of Karen's overall performance."	Memory Instructions + "Please give a rating of the Karen's performance on memory."
Screen 2	"Please give a rating of Karen's overall performance."	"Please give a rating of the Karen's performance on memory."		Cooperation Instructions + "Please give a rating of Karen's performance on cooperation."
Screen 3		Cooperation Instructions + Explicit Instructions		
Screen 4		"Please give a rating of Karen's performance on cooperation."		

Note.

Explicit Instructions:	On the next screen you will be asked to give a rating of Karen's (Mike's) performance. Before making your rating, we want you to think analytically about the reasons for your rating. Put your emotions aside and type, in the space below, the reasons you think are important in providing your rating.
Overall Instructions:	We are interested in Karen's (Mike's) overall performance. How well did Karen perform all of her duties?
Memory Instructions:	We are interested in Karen's (Mike's) performance with respect to memory. How well did Karen remember orders, bring the correct orders to the guests, and remember the correct beverages and side orders?
Cooperation Instructions:	We are interested in Karen's (Mike's) performance with respect to cooperation. How well did Karen accept guidance from supervisors and coworkers, and cooperate with and help other servers, supervisors, and wait staff?

Table 6
Dependent measures summary

Measure	Description	Formula
Overall rating accuracy	Subtract the participant's overall rating from the mean overall expert rating. Take the absolute value.	abs(participant's overall rating - mean expert overall rating)
Memory rating accuracy	Subtract the participant's memory rating from the mean memory expert rating. Take the absolute value.	abs(participant's cooperation rating - mean expert cooperation rating)
Cooperation rating accuracy	Subtract the participant's cooperation rating from the mean cooperation expert rating. Take the absolute value. .	abs(participant's memory rating - mean expert memory rating)
Specific rating accuracy	Take the mean of the cooperation accuracy rating and the recognition accuracy rating.	mean(accuracy of memory rating, accuracy of overall rating)
Rating accuracy	In the overall condition, rating accuracy is equal to the overall rating accuracy. In the specific condition, rating accuracy is equal to the specific rating accuracy (the average of the cooperation and memory rating accuracy).	If in overall condition: overall rating accuracy. If in specific condition: specific rating accuracy
Recognition accuracy	Ability to distinguish signal from noise. Approximately equal to the percentage of memory questions answered correctly.	NORMDIST((NORMSINV(Hit rate) - NORMSINV(False alarm rate))/SQRT(2))

Note. abs = absolute value.

Table 7

Mean performance appraisal rating as a function of condition

		Low Load				High Load			
Rating	Justification	N	Mean	SD	True Score	N	Mean	SD	True Score
Cooperation	Implicit	39	5.64	1.14	7.00	42	6.02	0.98	6.00
	Explicit	39	5.33	1.22	7.00	41	5.78	1.17	6.00
Memory	Implicit	39	2.26	0.97	1.25	42	2.67	1.30	3.08
	Explicit	39	2.67	0.90	1.25	41	3.10	1.26	3.08
Overall	Implicit	33	3.36	1.27	3.63	44	4.32	1.27	3.90
	Explicit	41	3.61	1.20	3.63	37	4.76	1.06	3.90

Table 8

Mean Performance Ratings

Rating	Judgment	N	Mean	SD	Mean Difference (Implicit vs. Explicit)
Cooperation	Implicit	81	5.84	1.07	0.28
	Explicit	80	5.56	1.21	
Memory	Implicit	81	2.47	1.16	0.42*
	Explicit	80	2.89	1.11	
Overall	Implicit	77	3.91	1.35	-0.24
	Explicit	78	4.15	1.27	

Note. * $p < .05$

Table 9
Mean rating accuracy and recognition accuracy scores

Rating	Justif-ication	Low Load					High Load				
		N	Rating Accuracy	SD	Recog-nition Accuracy	SD	N	Rating Accuracy	SD	Recog-nition Accuracy	SD
Specific	Implicit	39	1.23	0.64	0.82	0.13	42	0.92	0.53	0.66	0.15
	Explicit	39	1.57	0.74	0.82	0.12	41	0.85	0.56	0.68	0.13
Overall	Implicit	33	1.07	0.71	0.80	0.15	44	1.01	0.86	0.66	0.15
	Explicit	41	1.04	0.58	0.81	0.10	37	1.11	0.79	0.65	0.15

Note. Rating accuracy represents deviation from the expert score: Higher scores indicate less accuracy. Recognition accuracy is based on the $A_{d'}$ statistic: Higher scores indicate greater accuracy.

Table 10
Analysis of Variance for Performance Rating Accuracy

Source	*df*	*F*	*p*
Experience as a server	1	0.00	0.96
Server Sex	1	0.00	0.95
Participant Sex	1	10.71	0.00*
Justification	1	1.45	0.23
Cognitive Load	1	11.51	0.00*
Rating Type	1	1.99	0.16
Justification * Cognitive Load	1	0.56	0.45
Justification * Rating Type	1	0.74	0.39
Rating Type * Cognitive Load	1	14.96	0.00*
Error	306	(0.46)	

Note. Value enclosed in parenthesis represents mean square error.
* $p < .05$

Table 11
Mean Performance Accuracy Ratings as a Function of Cognitive Load and Rating Type

Cognitive Load	Rating Type	N	Mean	SD
Low	Specific	78	1.40	0.71
	Overall	74	1.05	0.64
	Specific and Overall	152	1.23	0.69
High	Specific	83	0.88	0.55
	Overall	81	1.06	0.83
	Specific and Overall	164	0.97	0.70
High and Low	Specific	161	1.13	0.68
	Overall	155	1.06	0.74

Table 12
Analysis of Variance for Recognition accuracy

Source	*df*	*F*	*p*
Experience as a server	1	0.41	0.52
Server Sex	1	7.70	0.01*
Participant Sex	1	4.84	0.03*
Justification	1	0.04	0.85
Cognitive Load	1	96.57	0.00*
Rating Type	1	1.16	0.28
Justification * Cognitive Load	1	0.04	0.84
Justification * Rating Type	1	0.03	0.87
Rating Type * Cognitive Load	1	0.07	0.80
Error	306	(0.02)	

Note. Value enclosed in parenthesis represents mean square error.
* $p < .05$

Table 13
Mean Ratings as a Function of Sex

Participant sex	Server Sex	N	Rating Accuracy Mean	Rating Accuracy SD	Recognition accuracy Mean	Recognition accuracy SD
Male	Male	54	1.30	0.77	0.71	0.15
	Female	59	1.15	0.75	0.71	0.17
	Both servers	113	1.22	0.76	0.71	0.16
Female	Male	100	0.98	0.60	0.71	0.16
	Female	103	1.07	0.73	0.79	0.13
	Both servers	203	1.02	0.67	0.75	0.15
Male and Female	Male Server	154	1.09	0.68	0.71	0.15
	Female Server	162	1.10	0.74	0.76	0.15

Table 14
Mean Confidence Ratings

		Low Load			High Load		
Rating	Judgment	N	Mean	SD	N	Mean	SD
Cooperation	Implicit	39	6.05	0.97	42	6.31	0.75
	Explicit	39	5.77	1.04	41	6.15	0.73
Memory	Implicit	39	6.08	0.90	42	5.88	1.21
	Explicit	39	5.90	0.99	41	5.85	1.17
Overall	Implicit	33	5.64	1.17	44	5.91	1.07
	Explicit	41	5.85	0.79	37	5.54	1.04

Table 15
Analysis of Variance for Confidence in Performance Ratings

Source	*df*	*F*	*p*
Experience as a server	1	0.24	0.62
Server Sex	1	0.80	0.37
Participant Sex	1	0.00	0.99
Justification	1	1.14	0.29
Cognitive Load	1	0.27	0.60
Rating Type	1	5.63	0.02*
Justification * Cognitive Load	1	0.82	0.37
Justification * Rating Type	1	0.12	0.73
Rating Type * Cognitive Load	1	1.78	0.18
Error	306	(0.83)	

Note. Value enclosed in parenthesis represents mean square error.
* $p < .05$

Appendix A

Sample Performance Appraisal Video Script

LOW PERFORMANCE

Memory- Brings orders to wrong people.

Setting: Two people- one male, one female- waiting to receive main course. Conversing about friend in MBA program.

G1 (Female): I've heard they serve delicious veal marsala here. I'm really glad I ordered it.

G2(Male): Yeah, I've heard the Steak Diane is quite good. A buddy of mine was here last week, and he said it was fantastic!

G1(Fe): Well, I'm starving! I can't wait to get our order.

Waitress approaches table carrying a tray.

G2(Ma): I am hungry too. Ah, here's our waiter...

Waitress places a meal in front of the two individuals.

G1(Fe): Ah super! Miss, I'm sorry, but this isn't what I ordered. I ordered Veal Marsala.

G2(Ma):(Addressing waitress) This doesn’t look like Steak Diane it looks like London Broil, or something.

Waitress: (repeats the correct order, apologizes, and leaves to get the Steak Diane and Veal Marsala) And you had Steak Diane, and you had Veal Marsala, I must have brought the wrong order. I’m really sorry I will get your correct orders.

Conversation about "Alan" resumes between the man and woman seated at the table.

Appendix B

Performance dimensions.

Dinner bill:
How well did the server present the bill to the appropriate guest at the appropriate time, and ensure that the bill had been added correctly and can be read and understood by the guests?

Maintaining performance levels:
How well did the server maintain patience, composure and good service when under pressure from crowds, large parties, or when tired?

Memory:
How well did the server remember orders, bring the correct order to the guest, and remember the correct beverages and side orders?

Cooperation:
How well did the server accept guidance from supervisors and coworkers, and cooperate with and help other servers, supervisors, and wait staff?

Work Habits:
How consistently did the server wear appropriate clothing and arrive at work on time? Was the server well groomed, prepared for work, and flexible to accommodate company needs?

Menu familiarity:
How well did the server display a familiarity with the food prices, ingredients, quality, portion sizes, and daily specials?

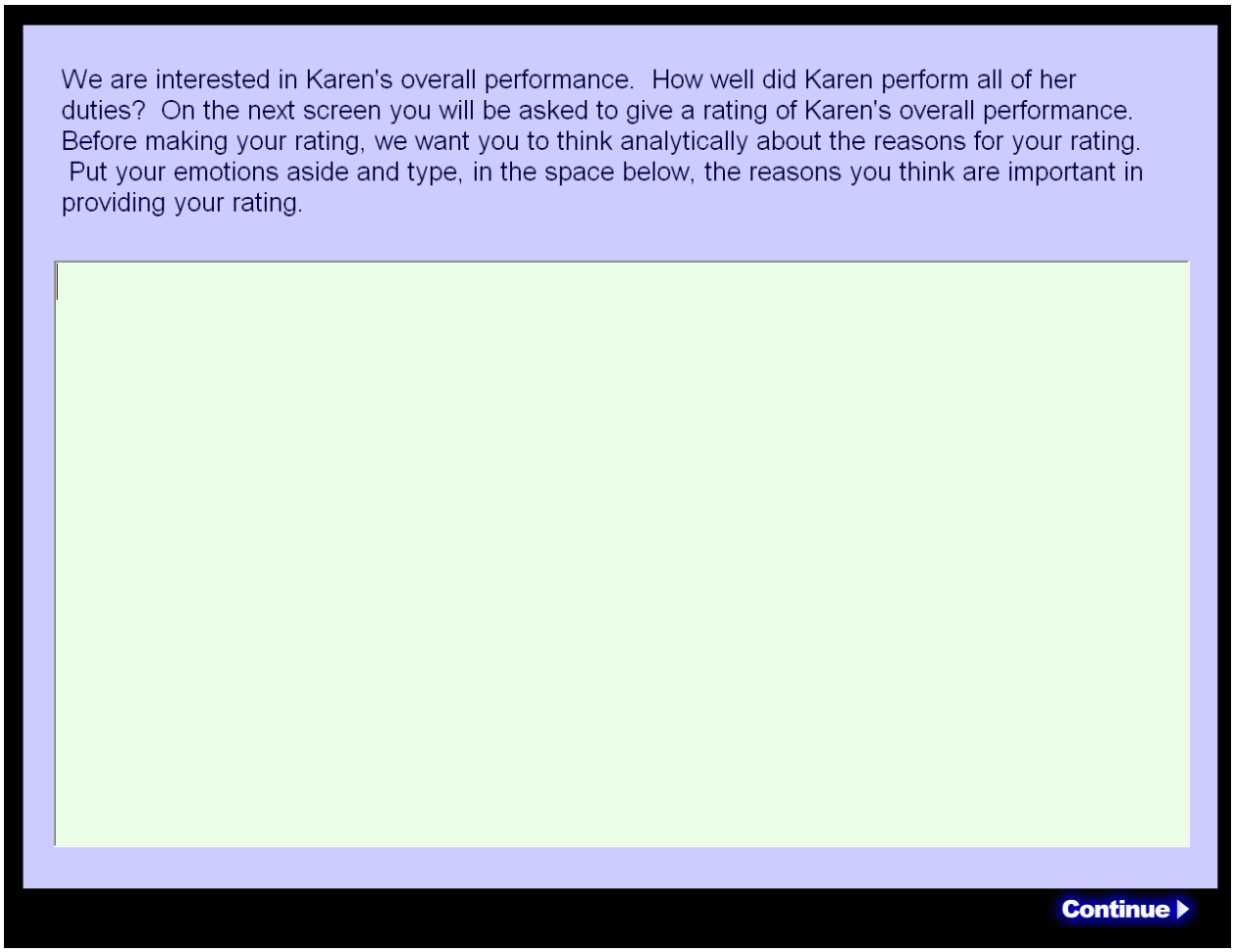
Appendix C

Opening Instructions

Thank you for coming in today. In this study you are going to be watching some videos of waiters and waitresses. You are then going to rate their performance. Please carefully read all of the instructions as you go through the study. This is a nice short study so you can take your time. Please let me know if you have any questions at any point. You can now read the instructions on the computer and begin.

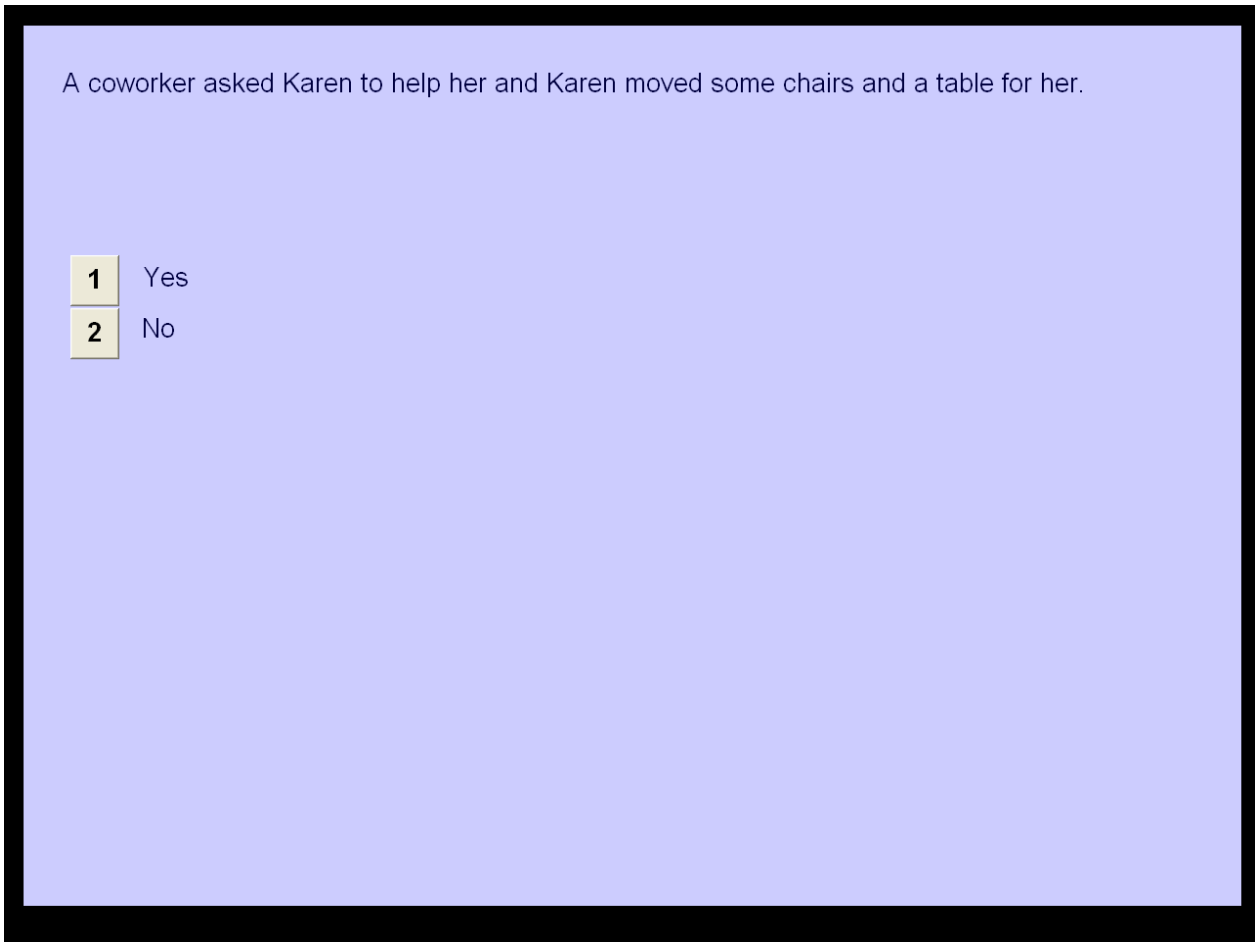
Appendix D

Explicit manipulation screenshot



Appendix E

Sample memory recognition question screenshot



Appendix F

Final rating choice