EFFECTS OF PERFORMANCE APPRAISAL PURPOSE AND RATER EXPERTISE ON RATING ERROR

by

WILLIAM S. WEYHRAUCH

B.S., Baker University, 2007

A THESIS

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Department of Psychology
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2010

Approved by:

Major Professor
Satoris S. Culbertson

Copyright

Abstract

Performance appraisals are an important component to any organization's performance management system. They require supervisors to observe and retain information regarding employee performance. This study sought to investigate the effects of appraisal purpose in this process. This extension and replication of Williams, DeNisi, Meglino, and Cafferty's (1986) lab study of appraisal purpose investigated whether designating an employee for a positive outcome results in lenient performance ratings and vice-versa for a negative designation. This outcome would indicate assimilation, whereby the designation acts as an anchor creating bias in the direction of the anchor. However, the negative and positive designations may both result in leniency, indicating a universal tendency toward leniency when memory for performance is limited. Furthermore, I investigated whether making a deservedness rating for each employee would result in less lenient or severe ratings, relative to the designation conditions. Finally, I investigated whether self-reported rater expertise would moderate the assimilation effect. A total of 108 undergraduate students from a large Midwestern university viewed confederates performing cardio-pulmonary resuscitation (CPR) on a dummy and were instructed to observe performance in order to make a designation (positive or negative) or deservedness rating, or were given no instructions (control). They made an initial decision and were then asked to return two days later and rate each confederate's performance again. Consistent with previous findings, raters making positive designations tended to give lenient ratings, relative to other conditions. Furthermore, as expected, those making negative designations gave relatively severe ratings. Finally, the results also partially supported my expectation that rater expertise in the performance domain moderates the biasing effects of appraisal purpose. Implications for practice and recommendations for future research are discussed.

Table of Contents

List of Figures

List of Tables

Acknowledgements

There are a number of people to thank for their guidance and assistance on this project. First and foremost, my thanks go to Dr. Satoris Culbertson for her guidance at every step of the way, including the conception, design, execution, and reporting of this study. Her mentoring over the past three years has meant so much to my development as a psychologist and as a person. Next, I want to thank the members of my committee, Drs. Patrick Knight and Jim Shanteau for their advice in improving the design and rationale. Thank you to my office mate Angela Connell for her assistance during research meetings and our daily commiserations. Furthermore, my thanks go to the undergraduate members of our research lab, Jamie Parker and Kevin Bowen for their substantial contributions to overcoming roadblocks in the execution of this and other studies. Last, but not least, I want to thank my helpful research confederates for their time and effort put into recording the performance stimuli videos used in this study: Katie Yunghans, Annie Weyhrauch, Kristin Babcock, and Jenny Parker.

Dedication

I dedicate this work to my family: to Mom, Dad, Katie, Annie, and the extended Weyhrauch and Robb families. The love and support I have received from all of you and the pride I take in calling you my family have sustained me at all times, especially the most challenging ones. The course of my life and education would have been very different without your patience, humor, inspiration, and love. Thank you all. To Mom, I love you very much and miss you every day. Thank you for your inspiring life and memory.

CHAPTER 1 – Literature Review

Performance appraisal is an organizational process within the umbrella of performance management in which an employee's job performance is evaluated and recorded. Performance appraisal information is used for a variety of organizational purposes. A common purpose is to contribute to such decisions as salary, promotion, and layoff decisions. In the performance appraisal literature, this is known as an administrative purpose (Cleveland, Murphy, & Williams, 1989; Jawahar & Williams, 1997; Landy & Farr, 1983; Levine, 1986). Administrative purposes can be further categorized into deservedness and designation rating purposes, a particularly important distinction when supervisors evaluate their employees in relation to an outcome (either positive or negative). A deservedness rating purpose occurs when employees are evaluated individually in relation to how worthy they are of (i.e., how much they deserve) a specific outcome. Conversely, a designation purpose occurs when supervisors attempt to identify (or designate) the single best or worst candidate for an outcome.

In addition to administrative purposes, performance appraisals can be used for developmental (e.g., feedback to help improve employee performance), systems maintenance (e.g., evaluations geared toward identifying training needs), and research-oriented (e.g., instrument validation) purposes (Boswell & Boudreau, 2002; Cleveland, Mohammed, Skattebo, & Sin, 2003; Cleveland et al., 1989). The distinction between these different performance appraisal purposes is important, as researchers have shown that the purpose of a supervisor's appraisal can affect how the performance information they observe is processed and subsequently judged (Bernardin & Beatty, 1984; Jawahar & Williams, 1997; Murphy, Balzer, Kellam, & Armstrong, 1984), a topic addressed in the present study.

When observing employee performance, supervisors may be considering multiple or all purposes for the appraisal. As Cleveland et al. (1989) noted, managers often use appraisal data to simultaneously make a variety of performance decisions, including between- and within-individual judgments. However, they may also consider only one use for their appraisal, when there could be many. For example, a manager may be faced annually with the need to decide which employees should receive a merit bonus (an administrative purpose) and will observe their performance with this in mind. However, the same supervisor may be asked later to evaluate the same employees in order to provide feedback to them (a developmental purpose). If appraisal purpose can affect the processing and storage of performance information, the nature of this processing and storage may influence ratings made subsequently for a different purpose. For example, if a supervisor has to identify one worker to be fired (or promoted), this might lead to negative (or positive) thought processes about each employee and lead to overly severe (or lenient) ratings for everyone on subsequent performance reviews. Leniency and severity are types of rating error in which a rater tends to give higher or lower (respectively) performance ratings than an employee deserves.

If, in fact, a supervisor's memory for performance can be influenced by the context present at the time performance is observed, it is worthwhile to investigate what processes are involved in this phenomenon and what factors may prevent or reduce the impact of these contextual variables on the quality of performance ratings. The present research study sought to investigate this issue and to provide insight on ways to address it. More specifically, I examined the effect of designation versus deservedness appraisal purposes on subsequent performance evaluations, the influence of outcome valence (i.e. positive or negative) on leniency/severity in

ratings, and the moderating effect of performance domain expertise on these leniency/severity effects.

Williams, DeNisi, Meglino, and Cafferty (1986) compared performance ratings made in the context of a designation or deservedness appraisal purpose to investigate how this difference in appraisal purpose affected subsequent overall performance ratings two days later. Their research revealed that when raters were told to view performance with the intent of immediately designating one worker for a positive outcome, as opposed to rating with a deservedness purpose, they rated all workers higher (i.e., rated more leniently) in subsequent performance ratings. They also found that raters with a deservedness purpose were better than raters with a designation purpose at differentiating the levels of worker performance in their ratings. The authors concluded that these differences were a result of a difference in the encoding taking place when the raters were observing performance with their assigned purpose in mind.

Williams et al. (1986) contributed to the understanding of the cognitive processes of raters conducting performance appraisals by supporting the preferability of person-structured observation (as opposed to task-structured) and also uncovering a possible assimilation effect when appraisals are structured around identifying a single worker as best. However, their findings and conclusions can be confirmed and extended in several ways, which is the purpose of this study.

First, the current study serves as a replication of Williams et al.'s (1986) research with a different conception of performance. Whereas Williams et al. had raters evaluate workers performing simple construction tasks (e.g., safely hammering a nail into wood), the present study used cardio-pulmonary resuscitation (CPR). This change was made to provide evidence for the generalizability of the findings, given the different performance context. Second, rater expertise

was incorporated into the current study to evaluate its potential to moderate the effects of appraisal purpose. Finally, an additional rating purpose, negative designation, was introduced.

Whereas Williams et al. (1986) examined the designation of an individual for a positive outcome (outside contracting work), they did not examine the designation of an individual for a negative outcome. The addition of a negative designation purpose allows for the investigation of whether the leniency observed by Williams et al. can be attributed to an assimilation effect, or a universal leniency effect in administrative ratings. An assimilation effect arises when an initial rating leads subsequent ratings to be biased towards the initial rating. Williams et al. speculated that a similar effect would be found in the case of a negative designation purpose, but resulting in rating severity instead of leniency, indicating an assimilation effect. However, their results do not rule out the universal leniency explanation.

Appraisal Purpose

Research on appraisal purpose has generally identified the following four categories of purposes: administrative, developmental, systems maintenance, and research (Greguras, Robie, Schleicher, & Goff, 2003; Harris, Smith, & Champagne, 1995; Jawahar & Williams, 1997). Appraisals with an administrative purpose are conducted to obtain performance information that will be used to make decisions related to promotions, pay, or other rewards and/or sanctions. Developmental appraisals are conducted to provide performance information that can be communicated back to the employee as feedback that will hopefully help them to improve their performance in the future.

Appraisals conducted for systems maintenance are intended to provide information on personnel planning and organizational training needs. Finally, appraisals for research purposes are typically not conducted for organizational purposes at all. Instead, in this case, performance

information is obtained in order to validate an instrument or contribute to an experimental research study.

As mentioned earlier, a more specific class of administrative appraisal purpose can be defined along the dimension of designation versus deservedness (Williams, DeNisi, Blencoe, & Cafferty, 1985). A designation purpose is one in which performance is observed with the intent of choosing a single worker for an outcome, either positive or negative. For example, a supervisor may be faced with choosing one employee to go through management training, a positive outcome that may lead to a promotion for that employee. A negative designation might be required when a supervisor is faced with budget cuts and must identify one employee to be laid off. Conversely, a deservedness purpose is characterized by providing a rating for each employee on whether or not they deserve a particular outcome, such as the ones listed above.

Appraisal purpose is an important element to understanding performance appraisal, due to the numerous ways by which it can unduly influence ratings. The negative effect of rating purpose on rating quality can be intentional or unintentional. Rating purpose may lead raters to intentionally rate inaccurately to achieve some other purpose, such as rating severely to motivate employees to work harder (Murphy & Cleveland, 1995) or rating leniently in order to reflect more kindly on their own management skills (Jawahar & Williams, 1997). Additionally, different purposes may place a variety of cognitive demands on the rater, leading to unintentional problems with ratings (DeNisi, Cafferty, & Meglino, 1984; Murphy & Cleveland, 1995). Different purposes may require raters to think about and evaluate performance in different ways. An administrative purpose, for instance, would likely lead raters to adopt a normative frame-of-reference that is relative to the performance of others being rated, whereas a developmental purpose would lead to an absolute frame-of-reference that is relative to some predetermined

standard (Jelley & Goffin, 2001). Researchers have found that administrative ratings of subordinates are significantly more lenient and less accurate than ratings for developmental and research purposes (Greguras et al., 2003). Therefore, it is worth investigating exactly which cognitive processes drive the reduction in quality seen in administrative ratings, such as the designation and deservedness ratings examined herein.

Information Processing

Spicer and Ahmad (2006) summarized several different cognitive process models proposed since 1980 that attempt to explain the stages of information processing used in performance appraisal (e.g., DeNisi et al., 1984; Feldman, 1981; Landy & Farr, 1980; Landy & Farr, 1983; Murphy & Cleveland, 1995; Wofford & Goodwin, 1990). While various stages have been added, modified, and removed in the more recent models, each model incorporates at least four basic stages: observation/acquisition of information, encoding/storage of information, retrieving information from memory, and evaluating the information as a whole to come to a judgment. The way in which information is structured by the mind when it is first presented (encoding) largely determines the way in which it is stored and later retrieved (Day & Sulsky, 1995; Hernstein, Carroll, & Hayes, 1980; Srull, 1983). Therefore, the cognitive processes of encoding have a significant impact on the accuracy with which information is recalled.

According to the process model described by Landy and Farr (1983), performance appraisal for any one employee begins with obtaining performance information. This commonly means a supervisor directly observing employee performance, as well as reviewing performance records and obtaining input from others. The information obtained during observation must be evaluated, either at the moment it is observed or later, drawing upon memory. Formal performance appraisals would not be adequately representative if performance was always

evaluated only on the basis of single performance events. This would be sampling only maximum performance, not typical performance. Maximum performance refers to a level of performance depicting what a worker *can* do during short, evaluative sessions, while typical performance reflects what a worker *will* do on a daily basis (Sackett, Zedeck, & Fogli, 1988). For example, if a supervisor observes performance once every appraisal period, and employees realize this is occurring, they will likely exhibit maximum performance levels at observation time, which may or may not be what the supervisor intends to evaluate. Therefore, to some extent, remembering performance information for later evaluation is always necessary.

A number of theories exist suggesting ways in which encoding strategy influences retrieval. The levels-of-processing framework (Craik & Lockhart, 1972) suggests that more meaningful information is likely to be remembered more clearly because it occurs at a deeper level of processing. Additionally, Tulving's (1983) encoding specificity principle proposes that memory is best when the conditions of retrieval match the conditions under which memories were encoded. A shared assertion of these theories is that the encoding context affects retrieval quality. This relates to the present research on appraisal purpose, which is conceptualized as a contextual factor present at the time of encoding. The present research seeks to provide further empirical evidence of the contextual processes that impact the encoding of performance information during performance observation.

According to the encoding specificity principle (Tulving, 1983), raters observing performance for a deservedness rating purpose should be more prepared to make subsequent ratings of each target's performance, as the context for their performance observation more closely matches this type of subsequent rating and so their memory of performance should be better. The levels-of-processing framework (Craik & Lockhart, 1972) provides further support

that performance observations made with a deservedness purpose will lead to greater memory for performance information because making individual deservedness ratings for each employee requires deeper levels of information processing than a single designation. In this case, greater memory is expected to lead to greater differentiation of worker performance, meaning that raters with greater memory for performance will provide significantly different ratings for workers at different levels of overall task proficiency. This expectation is based on the findings of Williams et al. as well as the theories described above, leading to the first hypothesis.

Hypothesis 1

Raters making designation decisions will be less able to differentiate worker performance than those making deservedness decisions.

Effects of previous decisions

In the present study, the context under investigation is a decision context, with an investigation of the memory effects of making a designation decision or a deservedness rating decision. Previous research has established that prior decisions/judgments can exert an enormous influence on subsequent ones (Lingle & Ostrom, 1979; Murphy, Balzer, Lockhart, & Eisenman, 1985; Smither, Reilly, & Buda, 1988; Thorsteinson, Breier, Atwell, Hamilton, & Privette, 2008). The influence of previous decisions is of particular interest to performance appraisal researchers, given that repeated, but ideally independent, judgments are routine and accompanied by considerably high stakes (e.g., the status of an employee's job security, salary).

In particular, this line of research investigates the assimilation effect. Assimilation is a type of rating phenomenon in which the distribution of ratings a person gives is influenced by the introduction of an anchor (Murphy et al., 1985). Assimilation refers specifically to rating error in the direction of an established anchor (Murphy et al., 1985). For example, if a supervisor

conducting job interviews meets the first applicant who happens to do very well, the interviewer may look more kindly on the rest of the applicants than otherwise. Contrast effects, on the other hand, refer to rating error in the opposite direction of an anchor (Murphy et al., 1985). In the example, the interviewer might view subsequent applicants more negatively than if the first applicant had not been so terrific. These effects are also referred to as context effects (Kravitz & Balzer, 1992), referring to the influence of the context (anchor) on the distribution of ratings, independent of what is being rated. These effects are examined by comparing ratings of the same target under different rating conditions (Kravitz & Balzer, 1992).

One theoretical explanation of the assimilation effect is the priming hypothesis proposed by Collins and Quillian (1969). According to this hypothesis, cognitive categories (such as positive, effective performance) used to organize the perception of one worker will prime the use of these categories in the perception of subsequent workers. In essence, thinking of an initial worker in terms of effective performance will produce benefits for subsequent workers by priming the rater to think in terms of effective performance behaviors.

In their research, Williams et al. (1986) found what appears to be an assimilation effect that might be explained by the priming hypothesis (although they did not refer to it this way). One of their central findings was that raters who were given a positive designation purpose subsequently gave more lenient ratings overall than raters who had a deservedness rating purpose. Their explanation of this finding was that the designation purpose limited the amount of performance information retained in memory for each worker because it did not require as much processing as the deservedness purpose. The designation purpose required less processing because there was no need to differentiate all levels of proficiency, just the best from the rest;

whereas, the deservedness purpose forced raters to evaluate each worker's individual performance.

However, less processing leads to a lack of memory for performance that may have a variety of effects on ratings, depending on the cause. If the leniency was caused by assimilation via the priming hypothesis, then a negative designation purpose would lead to severity in ratings. Conversely, it may be that the lack of memory for performance results in leniency. Researchers have documented that administrative ratings (including both deservedness and designation ratings) do tend to be more lenient (Jawahar & Williams, 1997), less variable (Fahr, Cannella, & Bedeian, 1991), and less accurate (McIntyre, Smith, & Hassett, 1984) than developmental ratings. These problems with administrative ratings may be exacerbated by the lower processing requirements of a designation purpose.

The present research design, which includes positive *and* negative designation conditions, allows for an examination of whether Williams et al.'s (1986) finding was a result of people's tendency to be lenient or if it was a result of assimilation. The assimilation explanation is derived from conceiving of the positive designation purpose as a positive anchor, priming raters to think in positive terms, thus leading to an evaluation of each worker with a positive frame of reference that may result in leniency. Despite the evidence of leniency in administrative ratings, the assimilation/priming hypothesis has a stronger theoretical basis, so Hypotheses 2 and 3 reflect the assimilation effect explanation.

Hypothesis 2

Raters with a positive designation purpose will demonstrate greater leniency in their ratings compared to those with a deservedness purpose.

Hypothesis 3

Raters with a negative designation purpose will demonstrate greater severity in their ratings compared to those with a deservedness purpose.

Experts and novices as raters

Horton and Mills (1984) noted that two general factors influence the encoding process: the organizational structure of stimulus material and the instructions for processing. In their research, Williams et al. (1986) conceived of the first factor, the organizational structure of stimulus material, as whether the information was presented in a format blocked by person (viewing one person performing all tasks, then moving onto the next person) or by task (viewing all workers performing each task, before moving on to the next task). They discovered that more information was retained in the instance of person-blocked formatting. The assigned appraisal purpose represented the second factor, the instructions for processing.

The present study investigates a different conception of Horton and Mills' (1984) organizational structure[1] and seeks to replicate and extend the findings of Williams et al. (1986) in relation to processing instructions through the addition of negative outcome designation and deservedness decisions. Instead of replicating the conditions of person- vs. task-blocking, I introduce performance domain expertise to investigate the organizational structure factor. The distinction between experts and novices is considered an example of organization of stimulus material in that experts, by definition (as discussed below), have sophisticated schemas and scripts that allow them to impose structure on performance information as it is being viewed, by directing their attention on certain aspects of the environment at crucial times. Therefore, from the very beginning, individuals with greater expertise should perceive a more organized set of performance information than novices.

Psychologists have asserted that the acquisition of expertise comes through the learning of domain-specific knowledge (Ericsson & Charness, 1994; Glaser, 1984). Among a variety of traits and processes used to characterize expertise are: (a) having large schemas rich with declarative knowledge about a given domain, (b) developing sophisticated representations based on structural similarities among problems, and (c) having schemas that contain procedural knowledge about strategies for solving a problem (Sternberg, 1998). The uniting theme of these three characteristics is the notion of a subject matter map (e.g., schema, script, representation) that experts use to guide their attention toward solving the problem. In performance appraisal, identifying mistakes in the performance of others may be seen as the problem. Thus, if a rater has a wealth of knowledge and experience in the form of schemas to rely on, he or she will likely have an easier time identifying mistakes in task performance.

Expert raters are conceptualized here as individuals who report high levels of experience and comfort with the performance domain, not necessarily expertise in rating it. This is a deviation from the traditional definition of expert raters, by which raters are characterized as having considerable knowledge in the performance domain *and* the procedure of rating it (Feldman, 1985). Using Feldman's (1985) conception of expert raters, Weekley and Gier (1989) found that expert raters were able to rate with superior reliability and validity levels, compared to earlier studies (Borman, 1978) on the reliability and validity of performance ratings. The present research was limited to participants reporting familiarity with the performance domain, not expertise in *rating* performance of others in that domain.

This alternative method of defining expertise examines a common hiring practice in which applicants for jobs higher up in an organization are recruited and selected from within the organization (internal labor markets; see Doeringer & Piore, 1985). A common result of a

promote-from-within selection procedure involves former job holders (incumbents) being promoted to positions of supervisory authority over their former position. However, given their experience with the job they now oversee, a new supervisor may not receive any explicit training on how to rate performance in that job. Novices are conceptualized as those with no (or very little) formal experience with the content of performance being rated. Experts should be better equipped to direct their attention to relevant information in the performance environment by relying on their schemas for the situation. In other words, self-reported expertise is expected to buffer the appraisal purpose effect, which is Hypothesis 4.

Hypothesis 4

Expertise will moderate the effects of appraisal purpose, such that experts' ratings will not be as influenced toward severity or leniency as novices.

Contributions of the present study

The current study tests the assumption made explicit in the discussion of Williams et al. (1986) that "the same processes would occur in the opposite direction when individuals are selected for negative outcomes" (p. 194). By replicating and extending their study, the current study examines empirically whether assimilation effects or universal leniency are responsible for errors resulting from designation purposes. The research design employed in the current study makes it possible to determine whether raters commit severity errors under a negative designation purpose, which would indicate an assimilation effect. Conversely, if raters are still lenient to the other workers, this would indicate that the leniency is a result of a universal tendency to be lenient in administrative ratings when there is less performance information retained. Finally, the current study contributes to research on cognitive processes in performance

appraisal by testing the relationship between these effects and the rater's level of performance domain expertise.

CHAPTER 2 - Method

Participants

In order to identify an adequate sample size necessary to detect effects should they be present, a power analysis was performed using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) based on the effect sizes reported in Williams et al. (1986). The necessary sample size was calculated for a power of at least 0.8, with an alpha of 0.05. The G*Power analysis indicated a necessary final sample size of 108.

Participants were recruited from General Psychology and upper-level psychology courses at a Midwestern university. Some attrition occurred between sessions, so data collection continued until the necessary sample size (as determined by the power analysis) had been reached. Research participation credit and/or extra course credit were awarded to participants as incentive to participate.

Materials

The primary stimulus materials included two videos: a CPR training video and a performance stimulus video. The CPR training video was excerpted from a DVD-based training program on CPR. The training involved a very brief (8 min.) demonstration and explanation of the proper technique and steps in administering CPR to an adult and to an infant. The steps of CPR were broken into four tasks. Though there are many important steps to be taken in an emergency situation (e.g., calling 911), chest compressions and breaths are the crucial elements in keeping a person alive. For the sake of simplicity and consistency with the four tasks used by Williams et al., adult and infant chest compressions and rescue breaths were chosen as the most appropriate way to divide CPR into distinct tasks.

The performance stimulus video showed footage of four research confederates performing CPR on an adult and infant dummy. Four confederates were trained and instructed to perform their tasks at one of three levels of proficiency: three of four tasks performed correctly (75% proficiency), two of four tasks performed correctly (50% proficiency), and one of four tasks performed correctly (25% proficiency). The two confederates at the 50% proficiency level committed mistakes on different tasks. Performance failure on a task was operationalized as a confederate making a clear error on a task, although not necessarily doing everything possible wrong on that task. The gender of the confederates was restricted to females to avoid any gender-related rating bias. Also, they were each in their early-mid twenties and of similar body type. An additional issue in developing the performance videos was the issue of whether mistakes on certain tasks might be perceived as more serious than mistakes on others (for example, mistakes made on an infant might seem more severe). The most proficient worker made her mistake doing the adult compressions. One middle proficiency worker erred on the adult compressions and infant breaths. To balance across this proficiency level, the other middle proficiency worker made mistakes on the other tasks: adult breaths and infant compressions. So, both middle proficiency workers made one mistake on the adult and one on the infant. Finally, the lowest proficiency worker made a mistake on all tasks except for the infant compressions.

Procedure

Participants were randomly assigned to one of four experimental conditions: a negative outcome designation condition, positive outcome designation condition, deservedness rating condition, or a control condition. They arrived at the experiment site and signed in to receive a participation code used to match their ratings at Time 2. They were then asked to fill out a basic demographic questionnaire, which included questions about their experience with CPR and their

CPR certification status (see questionnaire in Appendix A). Once all participants had arrived and completed the required paperwork, the experimenter gave a brief summary of the purpose of the research. Participants were told that they were part of a study designed to select applicants to be in a future research project that would require a group of participants who are competent in CPR. They were also told that this study had been sponsored by a grant from the National Institutes of Health, which would allow those participants to be paid. The researcher emphasized that the pay was substantial and that ratings of the applicants should be taken seriously, as they would be used in determining who was accepted into that study (see researcher scripts in Appendices G-J). This deception was done in order to enhance ecological validity by creating experimental conditions that matched an administrative appraisal purpose, rather than a research purpose.

Participants then viewed the CPR training video and were also provided with a written summary of the most important aspects of proper CPR technique for each task (see Appendix B). The training video and written guidelines were used to provide participants with knowledge of what is and is not correct CPR[2]. Furthermore, watching the video was not expected to do much to alter the difference between novices, who only know as much as they learn from the video, and experts, who have been through an appropriately designed course involving hands on practice, repetition and direct instruction. At this point, participants were given varying types of instructions for how to observe the performance videos, depending on their appraisal purpose condition.

Conditions

Participants in the positive outcome designation condition were instructed both verbally and on their rating sheet to watch the video with the purpose of choosing one of the four confederates who, based on her performance, most deserved to be accepted into the study.

Participants in the negative outcome designation condition were instructed to watch the video with the intent of identifying one performer as the most deserving of being eliminated from the applicant pool. Those in the deservedness rating condition were told that their job was to rate each performer using a 7-point Likert scale indicating whether the applicant should be accepted into the study or rejected. The scale was anchored at 1) *Definitely should be eliminated from consideration for the NIH study*, 4) *Adequate performance, but not the best candidate*, and 7) *Would be an ideal candidate for the NIH study*. Finally, those participants assigned to the control condition viewed the video with no appraisal purpose provided to them beforehand. They only knew they would be expected to evaluate the study applicants after viewing their performance.

Participants then viewed the performance video. The video showed each confederate performing the four tasks. Each confederate was given a fake first name which appeared on the screen to identify her before her video began. To avoid order effects, four versions of the performance video were made, such that each confederate was shown in each position (i.e., first, second, third, or fourth) once. The four versions of the video were then randomized across sessions.

Afterward, participants in each condition made the appropriate appraisal decision/rating for their experimental condition. Those in the control condition were given the same rating form as those in the deservedness condition. Participants were dismissed and asked not to discuss the study with other participants and reminded to return for the second session two days later. Some limited attrition (~10%) occurred between sessions, but appeared to be random, not systematically related to any of the conditions.

Two days later, participants in all conditions returned to the same testing site and were asked to recall the performance of the confederates and make global and task-specific quality

ratings for each confederate's performance. The performance ratings were made on a 7-point Likert scale with anchors at 1) *Poor*, 4) *Average*, and 7) *Outstanding*. The rating sheet included a small photograph of each of the confederates to help participants recall each confederate (see Appendix F).

CHAPTER 3 - Results

The mean overall performance ratings made in Session 2 are presented in Table 1, organized by appraisal purpose condition. A 4 (appraisal purpose) x 3 (worker proficiency) mixed factorial ANOVA was conducted, treating worker proficiency as a within-subjects variable and appraisal purpose as a between subjects variable. A manipulation check and tests of Hypotheses 1-3 were tested in a single mixed factorial ANOVA. The manipulation of worker proficiency was checked by probing the main effect of worker proficiency in the ANOVA. Hypothesis 1 was tested by the interaction of appraisal purpose and worker proficiency, while Hypotheses 2 and 3 were tested in the main effect of appraisal purpose. Although ANOVAs are normally discussed in order of main effects, followed by interactions, in order to report the manipulation check and hypotheses in the same order presented in the introduction, the main effect of worker proficiency is reported first, then the interaction, and finally the main effect of appraisal purpose.

Before interpreting any results, the assumption of sphericity was checked. Mauchly's test of sphericity was significant, indicating a violation of this assumption. However, the results were robust to the violation, as the Greenhouse-Geisser, Huynh-Feldt, and Lower-bound corrections all resulted in the same F ratios. The main effect of appraisal purpose was significant, $F(3,104) = 14.344$, $p < .001$, $\omega^2 = 0.11$, as well as the main effect for worker proficiency, $F(2, 208) = 224.082$, $p < .001$, $\omega^2 = 0.58$. The interaction was also significant, $F(6, 208) = 3.662$, $p < .05$, $\omega^2 = 0.01$. Effect sizes were estimated using ω^2, a relatively unbiased estimate that is unaffected by sample size (Carroll & Nordholm, 1975).

The manipulation of worker proficiency was intended so that there would be three distinct levels of proficiency. Probing the main effect of worker proficiency level demonstrated

that raters perceived different levels of performance among the workers. Post-hoc comparisons using Tukey's HSD (q) statistic indicated ratings for the high proficiency (75%) worker (M = 5.27) were significantly higher than ratings for the medium proficiency (50%) workers (M = 3.89, $q(104) = 14.96$) and the low proficiency worker ($M = 2.69$, $q(104) = 13.04$). The low proficiency worker was also rated significantly lower than the medium proficiency workers. This provides evidence that the intended manipulation of worker proficiency level was effective.

No steps were included in the study procedures to conduct a manipulation check for the cover story, testing whether participants believed that their ratings were going to be used in selecting participants for a future research project. However, after debriefing at the end of Session 2, anecdotal evidence suggested that many were surprised to hear that the cover story was entirely fictional and that their ratings would not be used for that purpose.

Hypothesis 1 was intended to be a replication of the results of Williams et al. (1986). They found that raters with a deservedness purpose were better able to differentiate workers at all levels of proficiency than those with a designation purpose. In the current study, the significant interaction of appraisal condition and worker proficiency indicates that the appraisal condition raters were assigned to did influence their ability to differentiate workers at various levels of performance. The interaction was probed with simple effects analysis by examining the effect of worker proficiency within each appraisal condition. All post-hoc mean comparisons were calculated using Tukey's HSD ($\alpha = .05$).

In the deservedness condition, ratings of the highest proficiency worker ($M = 5.07$) were significantly higher than ratings of the medium proficiency workers ($M = 3.93$, $q(26) = 11.56$). Likewise, ratings of the medium proficiency worker were significantly higher than the lowest proficiency worker ($M = 3.19$, $q(26) = 7.50$). In the positive designation condition, ratings of the

highest proficiency worker (M = 5.89) were significantly higher than ratings of the medium proficiency worker (M = 4.52, q(26) = 14.88), which were in turn significantly higher than the lowest proficiency worker (M = 3.48, q(26) = 11.30). In the negative designation condition, ratings of the highest proficiency worker (M = 4.64) were significantly higher than ratings of the medium proficiency worker (M = 3.35, q(26) = 16.82), which were in turn significantly higher than the lowest proficiency worker (M = 2.07, q(26) = 16.69). Finally, in the control condition, ratings of the highest proficiency worker (M = 5.48) were significantly higher than ratings of the medium proficiency worker (M = 3.76, q(26) = 23.15), which were again significantly higher than the lowest proficiency worker (M = 2.04, q(26) = 23.15).

In sum, raters in all conditions were able to differentiate each level of worker performance. Therefore, Hypothesis 1 is not supported and the findings of Williams et al. (1986) were not replicated. A graph of these means is presented in Figure 1.

Hypotheses 2 and 3 consist of a replication and extension of the findings of Williams et al. (1986). Their results showed that raters with a positive designation purpose gave lenient ratings relative to raters with a deservedness purpose. This is Hypothesis 2 in the current study. Additionally, I predicted that a negative designation purpose would result in severe ratings, relative to raters with a deservedness purpose. To test these hypotheses, the significant main effect of appraisal purpose was probed (again with Tukey's HSD comparisons; α = .05) to identify differences in raters' Session 2 overall performance ratings between purpose conditions. Post-hoc comparisons revealed that raters in the positive designation condition (M = 4.63) gave significantly higher overall performance ratings (across proficiency levels) than raters in the negative designation (M = 3.35; q(26) = 18.02), deservedness (M = 4.06, q(26) = 8.02), and control (M = 3.76, q(26) = 12.25) conditions. Conversely, raters in the negative designation

condition (M = 3.35) gave significantly lower overall performance ratings than raters in the positive designation (M = 4.63, q(26) = -18.02) and deservedness conditions (M = 4.06, q(26) = -9.99), but not the control condition (M = 3.76, q(26) = -5.77). Finally, ratings made in the deservedness condition (M = 4.06) were significantly lower than ratings in the positive designation condition (M = 4.63, q(26) = -8.02) and significantly higher than ratings in the negative designation condition (M = 3.35, q(26) = 9.99). Mean ratings were not significantly different between the deservedness (M = 4.06) and control (M = 3.76, q(26) = 4.22) conditions. Thus, Hypotheses 2 and 3 were both supported.

A further check on Hypotheses 2 and 3 was conducted by analyzing the interaction again, but using simple effects analysis to break down the effect of appraisal condition, holding proficiency level constant. This analysis indicates to what extent the main effect of appraisal condition is consistent across the worker proficiency levels. All post-hoc comparisons were again calculated with Tukey's HSD tests (α = .05). These differences are summarized in both Figure 1 and the subscripts in Table 1.

The results indicate further support for Hypotheses 2 and 3 by demonstrating that the leniency/severity effect applied fairly consistently across worker proficiency levels, with the exception of the control condition.

Hypothesis 4 proposed a moderation effect of performance domain expertise, such that raters high in familiarity with CPR (experts) would provide ratings less influenced by appraisal purpose than those low in familiarity (novices). To test this, ratings of CPR familiarity were dichotomized into a new variable (CPR Expertise) with two levels: Expert and Novice. Experts and novices were operationally defined in two different ways. The first, a stricter definition, grouped and labeled participants who rated themselves a 1 or 2 on familiarity with CPR

(indicating low levels of familiarity) as novices. Likewise, those who rated themselves a 6 or 7 (indicating high levels of familiarity) were grouped and labeled as experts. This resulted in 29 cases identified as novices and 24 cases identified as experts. A 4 (appraisal purpose) x 2 (expertise) between-subjects multivariate ANOVA was performed on overall performance ratings of each worker proficiency level. The interaction of condition and expertise was significant for the medium, $F(3, 45) = 2.93$, $p < .05$, and lowest worker proficiency levels, $F(3, 45) = 12.95$, $p < .001$. The interaction, was not significant for the highest proficiency level, $F(3, 45) = .867$, *ns*.

The second definition of CPR familiarity was conducted using a wider definition of experts and novices. Subjects who rated themselves a 3 on familiarity were also labeled novices and those who rated themselves a 5 were included as experts, leaving only those who rated themselves a 4 (the midpoint of the scale) being excluded. This added 11 participants as novices and 18 cases as experts, resulting in a sample of 40 novices and 42 experts. The same analysis was conducted with the new definition of expertise and similar results emerged. The interaction was significant for ratings of the lowest proficiency level, $F(3,74) = 6.22$, $p = .001$, but not for the highest, $F(3,74) = 1.74$, *ns*, or the medium, $F(3,74) = .676$, *ns*, proficiency levels.

Figures 2-4 show the means for experts and novices for each condition and at each proficiency level. By examining the means plotted in Figures 2-4, a general pattern emerged that showed that the experts' ratings tended to be less influenced by the condition, particularly in the positive and negative designation conditions. This pattern is seen in the relative stability of the expert line relative to the novice line. The fact that each expert line is not perfectly horizontal suggests that the leniency/severity effects are still present. However, relative to the corresponding novice line, each expert line appears to have greater stability across conditions.

Although the interaction was only significant at the lowest level of proficiency, the general pattern is supportive of the hypothesized effect.

Figures 5-7 exhibit the same analysis as Figures 2-4, only using the looser definition of experts and novices. The pattern in these figures is similar, although the means for novices and experts appear slightly less stable across conditions, as would be expected given the more inclusive definition of expertise. In sum, because the interaction between condition and expertise was only significant for two worker proficiency levels with the strict definition and one worker proficiency level with the loose definition, Hypothesis 4 is, at best, partially supported.

CHAPTER 4 - Discussion

The purpose of this study was to expand on the results of Williams et al.'s study (1986) by answering some questions left unaddressed by their study design. Of primary interest were two questions: 1) would a negative designation condition result in severe ratings, instead of lenient ratings, and 2) does expertise in the performance domain reduce or eliminate rating error caused by appraisal purpose? The results of this study provide answers to these questions and also provide extra support for the importance of context and purpose in performance appraisal.

The findings of Williams et al. (1986) were partially confirmed and extended. Consistent with Williams et al.'s findings, raters assigned to observe performance and designate one worker for a positive outcome gave higher subsequent ratings than those given a deservedness purpose. However, the present study's findings did not replicate Williams et al.'s finding that raters with a deservedness purpose were better able to differentiate workers at different levels of proficiency than those with a designation purpose. On the contrary, the present study found that raters in all conditions were able to statistically differentiate between the best worker and the average workers, as well as between the average workers and the worst worker. Finally, consistent with the findings of Williams et al. (1986), the present study resulted in lenient ratings when raters were given a positive designation purpose, compared to those with a deservedness rating purpose.

Based on the results of Williams et al. (1986), as well as research in cognitive psychology, I hypothesized that raters in the deservedness condition would be able to differentiate the workers at different proficiency levels better than the raters in the designation conditions. This was expected to occur as a result of greater processing requirements in the deservedness condition. However, the expected superiority of raters in the deservedness

condition did not materialize. Rather all conditions were able to differentiate all levels of worker proficiency. The most obvious explanation for this difference is that the current performance videos may have made the errors more obvious. Furthermore, a fairly convincing story given to the participants about what their ratings would be used for combined with the more serious nature of CPR as a task (instead of woodworking) may have led the participants in this study to take their job more seriously and try harder to remember who was better than whom.

A key extension in the present study was the addition of a negative designation decision. As expected, and consistent with speculation put forward by Williams et al. (1986), the negative designation decision resulted in severe ratings, compared to those with a deservedness rating purpose. This provides evidence in favor of the assimilation explanation of both Williams et al.'s findings and those of the present study, rather than the universal leniency explanation. This should not be confused as proof of assimilation as the cause of these ratings. Rather, these findings disconfirm the universal leniency explanation and provide evidence for either assimilation or some other unidentified cause.

Another extension built into this design was the role of rater expertise in the performance domain. These findings indicate that people who considered themselves experts or novices in CPR had different patterns of rating error. Specifically, the ratings of experts were generally more resistant to the error caused by appraisal purpose. Although the interaction was not significant for all worker proficiency levels, the pattern is evident in the relative stability of expert ratings across conditions compared to novice ratings. Figures 2 and 5 exhibit the clearest examples of this relative stability. The brief DVD-based training video and written guidelines may have limited the impact of the expertise variable by shrinking the knowledge gap between people coming into the study with familiarity in CPR and those coming in with no experience.

Strengths and Limitations

This study has several strengths worth mentioning. This study employed a strong experimental design that permitted an exploration of two questions left unanswered by the work of Williams et al. (1986), namely the additional designation condition and the familiarity ratings used to assess rater expertise in the performance domain. Additionally, the inclusion of a control condition is also an extension of the Williams et al. study. This addition was based on the recommendation of Kravitz and Balzer (1992) who criticized the standard design used for studies on context effects in performance appraisal and recommended the use of a random context control condition to avoid ambiguous experimental results.

Furthermore, a variety of steps were taken to strengthen the internal validity of the study, including random assignment to conditions, counterbalancing the order of performance videos, using only females as confederates to avoid gender effects, using only females with similar body types to avoid bias related to physical attractiveness, and careful assignment of mistakes to each confederate so that issues of greater sympathy for infants would be at least partially resolved.

Despite the numerous strengths of this study, there are also some limitations that should be noted. The laboratory setting of this study limits the generalizability of the results. Williams et al. (1986) noted this limitation of their study and noted that a major issue was whether raters with expertise in the performance domain would have "developed scripts, or schemata, to aid them in their performance appraisals" (p. 194). The current study investigated this and discovered evidence supporting their speculation. This suggests that an expert who is very familiar with the performance domain of the job may have less to worry about when making performance appraisals in the context of a designation purpose. The generalizability of the study is still limited by the performance context. Although having replicated some of the results of Williams et al.

with a different conceptualization of performance, it remains to be seen how these results would apply to real-world appraisal situations where performance is evaluated over longer time periods by actual supervisors and involving more significant stakes.

Another limitation may be the nature of the mistakes made by the confederates in the performance stimulus videos. For example, mistakes made on the infant tasks may have been perceived as more egregious errors than mistakes on the adult tasks, due to the vulnerability and innocence of an infant. Steps were taken in the design of the study to deal with this issue. Specifically, the medium proficiency workers' mistakes were counterbalanced, with neither making an error on the same task and both making an error on an adult task and an infant task. Furthermore, the highest proficiency worker's one mistake was on an adult, not an infant. The lowest proficiency worker, on the other hand, got only one task right, which was the infant compressions task. The fact that the lowest proficiency worker's correct task was on an infant may have led to increased scores for the lowest proficiency level. In fact, across all conditions, ratings for the lowest proficiency worker were higher than the hypothetical true score of 1.75 (based on dividing the seven point scale by four).

Furthermore, there is some controversy among some health experts about whether rescue breaths should be eliminated from CPR guidelines because the lost time doing chest compressions outweighs the benefits (Ewy, 2007). This might have led participants who were aware of this controversy to weight mistakes on chest compressions more heavily than mistakes on rescue breaths. Participants in the current study were instructed to evaluate performance based on the guidelines provided, not their previous education or opinion. However, they may have ignored this instruction, either intentionally or unintentionally. Although this issue is specific to

this task, future researchers would be well-advised to consider the potential for differential weighting for certain tasks over others.

The conceptualization of expertise in this study may be considered a limitation. The issue is primarily one of terminology. Technically, because no test of knowledge was used to validate participants' self-reported familiarity with CPR, there is no evidence that they truly are experts or novices. Therefore, I refer to self-reported expertise to emphasize that participants were labeled as experts when they reported a sense of competence in their understanding of the performance domain. As it turned out, the results for the most part reflected the hypothesized moderation effect that was based on the notion that experts would be better able to identify errors and avoid rating bias.

A final limitation is the way in which the control condition was conceptualized. In retrospect, a more appropriate control condition might have involved no initial ratings at all in Session 1, instead of having these participants make the same ratings as the deservedness group. Alternatively, it may have been more appropriate to have a control condition for each type of designation, although this would have resulted in a cumbersome design involving four control conditions. Nevertheless, future researchers may wish to examine alternative control conditions in their examinations of these phenomena.

Implications for Research and Practice

The clearest implications of this study lie in the design of performance appraisal procedures. Procedures that structure the way supervisors observe employee performance in line with the purpose of subsequent appraisals should result in less biased ratings. For example, it would be beneficial to provide supervisors with a standardized form to guide their observations of performance with the same dimensions that will later be used in evaluation. Furthermore, the

results of the current study suggest that, even if performance appraisals are primarily used for designation purposes (e.g., identifying those with management potential), each employee should be evaluated equally on their deservedness for that outcome.

These implications, however, were present in the work of Williams et al. in 1986. The most significant implication of this new design is the assertion that, despite the widely accepted notion that administrative ratings will inevitably be lenient, when appraisal is conducted for the purposes of a negative outcome, rating error shifts toward severity. This finding is particularly relevant in the current economic climate. Layoffs due to budget constraints have forced many managers to identify workers to be terminated who might have kept their jobs in better circumstances. Although research on appraisal purpose has dwindled in the industrial psychology literature, this research demonstrates that not all our questions in this arena have been resolved.

Another unique contribution of the current study is the implication that supervisors who are very familiar with the jobs of their subordinates may be better equipped to observe and retain relevant performance information. Prior experience in the job they will oversee could be a criteria used in selection of supervisors from outside the organization. Conversely, organizations could provide training to give supervisors knowledge of the performance domain. There are a variety of ways to do this, such as providing supervisors with the same training given to new employees in the positions they oversee. Beyond that, it may be beneficial for supervisors to experience performing the job themselves for a period of time. A promote-from-within policy could help in this area. There is some controversy in the management literature regarding whether effective managers must be able to effectively do the work of employees they oversee or if some people are just good managers, but would not necessarily be good in the job they supervise. The findings of this study point out a weakness in the "good manager" position by

demonstrating that raters who have less knowledge of the performance domain are more prone to rating error.

The use of a CPR task in this study may make this research particularly interesting and useful to individuals involved in occupational health and/or CPR training. Individuals who design CPR training programs may be interested to see that the vulnerability of an infant, compared to an adult, may have some effect on the way CPR trainers evaluate a student's proficiency.

In regard to research implications, it appears that the effect of a deservedness purpose versus a designation purpose on the ability to differentiate all performance levels may not be a consistent effect. While it may have been an issue of the current sample or an artifact resulting from this conceptualization of performance, future research may be warranted on what may moderate the negative impact of designation on subsequent performance ratings.

There are several research implications based on the type of task performance used by this study and that of Williams et al. (1986). One point of difference between the CPR task and the Williams et al. woodworking task is that the result of a mistake in woodworking is a bad product, or possibly injury to the employee. The risk of a mistake in CPR is serious injury and/or death to an innocent person. This may have resulted in raters taking their task more seriously in the present study. Also, for the laboratory nature of this research, both the CPR task and the woodworking task were ideal because mistakes can be clearly defined and operationalized. However, on many jobs, errors are not so clearly defined and observable. So, it remains to be seen how using a more cognitive task with errors that are harder to observe would influence the results. A field study of the rating phenomena found in the current study would bolster the evidence of generalizability.

Occupational health psychologists may be interested in the difference between mistakes that necessarily lead to bad task outcomes and those that are violations of a rule created for the employee's safety. Examples include wearing goggles during woodworking or keeping elbows locked during CPR. Neither of these would necessarily lead to poor performance on the task, but may adversely affect the person doing it (e.g., keeping your elbows locked during CPR reduces the effort required to get the right amount of compression). One final task-related issue for researchers to address is how raters evaluate tasks that have multiple strategies for completion. Supervisors who endorse a particular strategy may have a hard time fairly evaluating an employee who uses an alternate, but equally valid, strategy. Thus, the choice of a task in this kind of performance appraisal research is a crucial element of study design and should be carefully considered in order to obtain the most generalizable results possible.

The study of rater expertise is an important area for future research. Instead of self-rated familiarity with the performance domain, future researchers may want to use an objective standard of expertise, or experimentally manipulate it by administering varying levels of training and/or examining the degradation of rater expertise in the performance domain over time. Furthermore, although the current study was focused on expertise in the performance domain, future research should investigate the potential mitigating effect of training in the rating process. This would permit a comparison of the appraisal purpose effect among expert raters in the traditional sense who have training in avoiding certain types of errors, raters with expertise in the performance domain who have highly developed schemas to help them identify errors, and those with both.

In addition to investigating the effect of different levels of rater training, future researchers could investigate the degradation of expertise over time. Introducing a time lag

between rater training and evaluating performance could result in unique rating effects. It may be that raters would more heavily weight the elements of performance that are more easily recognized as right or wrong, or perhaps they would remember the most crucial elements and rate primarily on these. In either case, this would more closely reflect the real-world conditions in which training is not as fresh in the mind of the rater.

Conclusion

In sum, the present study confirms the importance of appraisal context on the accuracy of performance appraisal ratings. By showing that raters who are evaluating with a positive or negative designation in mind for one worker may give lenient or severe ratings, respectively, I have demonstrated the importance of structuring performance appraisal procedures to maximize the amount and quality of information observed and retained about each employee's performance. One way to do this would be to provide raters with expertise in the performance domain, either through extensive training or through real experience doing the job of each employee they are assigned to evaluate. According to the current findings, this may provide raters with a greater ability to avoid the well-documented biases that can result from evaluating performance with a certain purpose in mind (DeNisi, Cafferty, & Meglino, 1984; Jawahar & Williams, 1997; Greguras et al., 2003; Murphy & Cleveland, 1995).

References

Bernardin, H. J., & Beatty, R. W. (1984). *Performance appraisal: Assessing human behavior at work.* Boston: Kent.

Borman, W.C. (1978). Exploring upper limits of reliability and validity in performance ratings. *Journal of Applied Psychology, 63*, 135-144.

Boswell, W.R. & Boudreau, J.W. (2002). Separating the developmental and evaluative performance appraisal uses. *Journal of Business and Psychology, 16*(3), 391-412.

Carroll, R.M. & Nordholm, L.A. (1975). Sampling characteristics of η2 and Hay's ω2. *Educational and Psychological Measurement, 35*, 541-554.

Cleveland, J. N., Mohammed, S., Skattebo, A. L., & Sin, H. P. (2003, April). Multiple purposes of performance appraisal: A replication and extension. Paper presented at the 18th annual meeting of the Society for Industrial and Organizational Psychology, Orlando, Florida.

Cleveland, J.N., Murphy, K.R., & Williams, R.E. (1989). Multiple uses of performance appraisal: Prevalence and Correlates. *Journal of Applied Psychology, 74*(1), 130-135.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Collins, A.M. & Quillian, M.R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning & Verbal Behavior, 8,* 240-248.

Craik, F.I. & Lockhart, R.S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning & Verbal Behavior, 11*, 671-684.

Day, D.V. & Sulsky, L.M. (1995). Effects of frame-of-reference training and information configuration on memory organization and rating accuracy. *Journal of Applied Psychology, 80*(1), 158-167.

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.

DeNisi, A.S., Cafferty, T.P., Meglino, B.M. (1984). A cognitive view of the performance appraisal process: A model and research propositions. *Organizational Behavior and Human Performance, 33*(3), 360-396.

DeNisi, A.S., Robbins, T., & Cafferty, T.P. (1989). The organization of information used for performance appraisals: The role of diary-keeping. *Journal of Applied Psychology, 74*, 124-129.

Doeringer, P. & Piore, M.J. (1985). *Internal Labor Markets and Manpower Analysis*. Armonk, NY: M.E. Sharpe.

Ericsson, K.A. & Charness, N. (1994). Expert performance: Its structure and acquisition. *American Psychologist, 49*(8),725-747.

Fahr, J., Cannella Jr., A.A., Bedeian, A.G. (1991). Peer ratings: The impact of purpose on rating quality and user acceptance. *Group and Organizational Studies, 16*, 367-386.

Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175-191.

Feldman, J.M. (1981). Beyond attribution theory: cognitive processes in performance appraisal. *Journal of Applied Psychology, 66*(2), 127-148.

Feldman, J.M. (1985). Instrumentation and training for performance appraisal: A perceptual-cognitive viewpoint. In K. Rowland & G. Ferris (Eds.), *Research in Personnel and Human Resources Management: Vol. 3*. Greenwich, CT: JAI Press.

Glaser, R. (1984). Education and thinking: The role of knowledge. *American Psychologist, 39*(2), 93-104.

Greguras, G.J., Robie, C., Schleicher, D.J., & Goff, M., III. (2003). A field study of the effects of rating purpose on the quality of multisource ratings. *Personnel Psychology, 56*, 1-21.

Hamilton, D.L., Katz, L.B., & Leirer, V.O. (1980). Cognitive processes in first impression formation. *Journal of Personality and Social Psychology, 39*, 1050-1063.

Harris, M.M., Smith, D.E., & Champagne, D. (1995). A field study of performance appraisal purpose: Research- versus administrative-based ratings. *Personnel Psychology, 48*, 151-160.

Hernstein, J. A., Carroll, J. S., & Hayes, J. R. (1980). The organization of knowledge about people and their attributes in long-term memory. *Representative Research in Social Psychology, 11,* 17-37.

Horton, D.L., & Mills, C.B. (1984). Human learning and memory. *Annual Review of Psychology, 35*, 361-394.

Jawahar, I.M. & Williams, C.R. (1997). Where all the children are above average: The performance appraisal purpose effect. *Personnel Psychology, 50*, 905-925.

Jelley, R.B. & Goffin, R.D. (2001). Can performance-feedback accuracy be improved? Effects of rater priming and rating-scale format on rating accuracy. *Journal of Applied Psychology, 86*(1), 134-144.

Kravitz, D.A. & Balzer, W.K. (1992). Context effects in performance appraisal: A methodological critique and empirical study. *Journal of Applied Psychology*, *77*(1), 24-31.

Landy, F.J., & Farr, J.L. (1980). Performance rating. *Psychological Bulletin, 87*(1), 72-107.

Landy, F. J., & Farr, J. L. (1983). *The measurement of work performance: Methods, theory, and applications.* New York: Academic Press.

Landy, F.J. & Farr, J.L. (1983). Cognitive aspects of the process model of performance rating: Theory and research. In F. Landy & J. Farr (Eds.). *The measurement of work performance; Methods, theory, and applications* (pp. 217-245. San Diego: Academic Press.

Levine, H. Z. (1986). Performance appraisals at work. *Personnel, 63*(6)*,* 63-71.

Lingle, J.H. & Ostrom, T.M. (1979). Retrieval selectivity in memory-based impression judgments. *Journal of Personality and Social Psychology, 37*(2), 180-194.

McIntyre, R.M., Smith, D.E., Hassett, C.E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. *Journal of Applied Psychology, 69*, 147-156.

Murphy, K. R., Balzer, W. K., Kellam, K., & Armstrong, J. (1984). Effect and purpose of rating on accuracy in observing teacher behavior and evaluating teaching performance. *Journal of Educational Psychology, 76,* 45–54.

Murphy, K.R., Balzer, W.K., Lockhart, M.C., & Eisenman, E.J. (1985). Effects of previous performance on evaluations of present performance. *Journal of Applied Psychology, 70*, 72-84.

Murphy, K.R. & Cleveland, J.N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives*. Thousand Oaks, CA: Sage.

Ostrom, T.M., Pryor, J.B., & Simpson, D.D. (1981). The organization of social information. In E. Higgins, C. Herman, & M. Zanna (Eds.), *Social cognition: The Ontario Symposium on Personality and Social Psychology* (pp. 214-265). Hillsdale, NJ: Erlbaum.

Sackett, P.R., Zedeck, S. & Fogli, L. (1988). Relations of typical and maximum job performance. *Journal of Applied Psychology, 73*(3), 482-486.

Sherman, S.J., Judd, C.M., & Park, B. (1986). Social cognition. *Annual Review of Psychology, 40*, 281-326.

Smither, J.W., Reilly, R.R., & Buda, R. (1988). Effect of prior performance information on ratings of present performance: Contrast versus assimilation revisited. *Journal of Applied Psychology, 73*, 487-496.

Spicer, D.P. & Ahmad, R. (2006). Cognitive processing models in performance appraisal: Evidence from the Malaysian education system. Human Resource Management Journal, *16*(2), 214-230.

Srull, T. K.. (1983). Organizational and retrieval processes in person memory: An examination of processing objectives, presentation format, and the possible role of self generated retrieval cues. *Journal of Personality and Social Psychology, 44,* 1157-1170.

Sternberg, R.J. (1998). Abilities are forms of developing expertise. *Educational Researcher, 27*, 11-20.

Thorsteinson, T.J., Breier, J., Atwell, A., Hamilton, C., & Privette, M. (2008). Anchoring effects on performance judgments. *Organizational Behavior and Human Decision Processes, 107*, 29-40.

Tulving, E. (1983). *Elements of episodic memory*. Oxford: Oxford University Press.

Weekley, J.A. & Gier, J.A. (1989). Ceilings in the reliability and validity of performance ratings: The case of expert raters. *The Academy of Management Journal, 32*(1), 213-222.

Williams, K.J., DeNisi, A.S., Blencoe, A.G., & Cafferty, T.P. (1985). The role of appraisal purpose: Effects of purpose on information acquisition and utilization. *Organizational Behavior and Human Decision Processes, 35*, 314-339.

Williams, K.J., DeNisi, A.S., Meglino, B.M., Cafferty, T.P. (1986). Initial decisions and subsequent performance ratings. *Journal of Applied Psychology, 71*(2), 189-195.

Wofford, J.C. & Goodwin, V.L. (1990). Effects of feedback on cognitive processing and choice of decision style. *Journal of Applied Psychology, 75*(6), 603-612.

Endnotes

[1] Instead of performing a true replication of Williams et al. (1986), the present study eliminates the person- and task-blocked conditions in favor of examining rater expertise. Research has supported and clearly shown that person-blocked processing is not only preferable (DeNisi, Robbins, & Cafferty, 1989; Williams, DeNisi, Blencoe, & Cafferty, 1985), but leads to greater accuracy and information retention compared to raters using task-blocked processing (DeNisi & Peters, 1996; Williams et al., 1986). Research in social cognition also supports the finding that person-blocked processing leads to more accurate judgments (Hamilton, Katz, & Leirer, 1980; Ostrom, Pryor, & Simpson, 1981; Sherman, Judd, & Park, 1986).

[2] It was emphasized that this does not constitute adequate CPR training and participants should seek out formal training before ever attempting to perform CPR.

Table 1: Mean overall performance ratings as a function of appraisal purpose

Worker proficiency (% correct)	Deservedness	Positive Designation	Negative Designation	Control	Total
75%	5.07_{ab}	5.89_{c}	4.64_{a}	5.48_{bc}	5.27
50%	3.93_{abc}	4.52_{a}	3.35_{bd}	3.76_{cd}	3.89
25%	3.19_{a}	3.48_{a}	2.07_{b}	2.04_{b}	2.69
Total	4.06	4.63	3.35	3.76	3.95

$N = 108$

Note: Within rows, means with common subscripts are not significantly different. Based on a 7-point rating scale, hypothetical true scores would be 5.25 (75% proficiency), 3.5 (50% proficiency), and 1.74 (25% proficiency). These true scores provide some indication of which mean ratings may be lenient/severe in reference to an absolute standard.

Table 2: Mean overall performance ratings as a function of appraisal purpose and expertise (using the strict definition of expertise)

Worker proficiency	Deservedness		Positive Designation		Negative Designation		Control	
	Expert	Novice	Expert	Novice	Expert	Novice	Expert	Novice
75%	5.20	4.50	5.60	6.00	4.83	4.43	5.63	5.33
50%	3.80	3.92	3.60	4.85	3.58	3.14	3.44	3.92
25%	1.80	3.67	1.40	4.30	2.25	2.17	1.25	2.50
Total	3.60	4.03	3.53	5.05	3.55	3.25	3.44	3.92

Note: Total $N = 108$, Expert $N = 24$, Novice $N = 29$. Based on a 7-point rating scale, hypothetical true scores are 5.25 (75% proficiency), 3.5 (50% proficiency), and 1.75 (25% proficiency). These true scores provide some indication of which mean ratings may be lenient/severe in reference to an absolute standard.

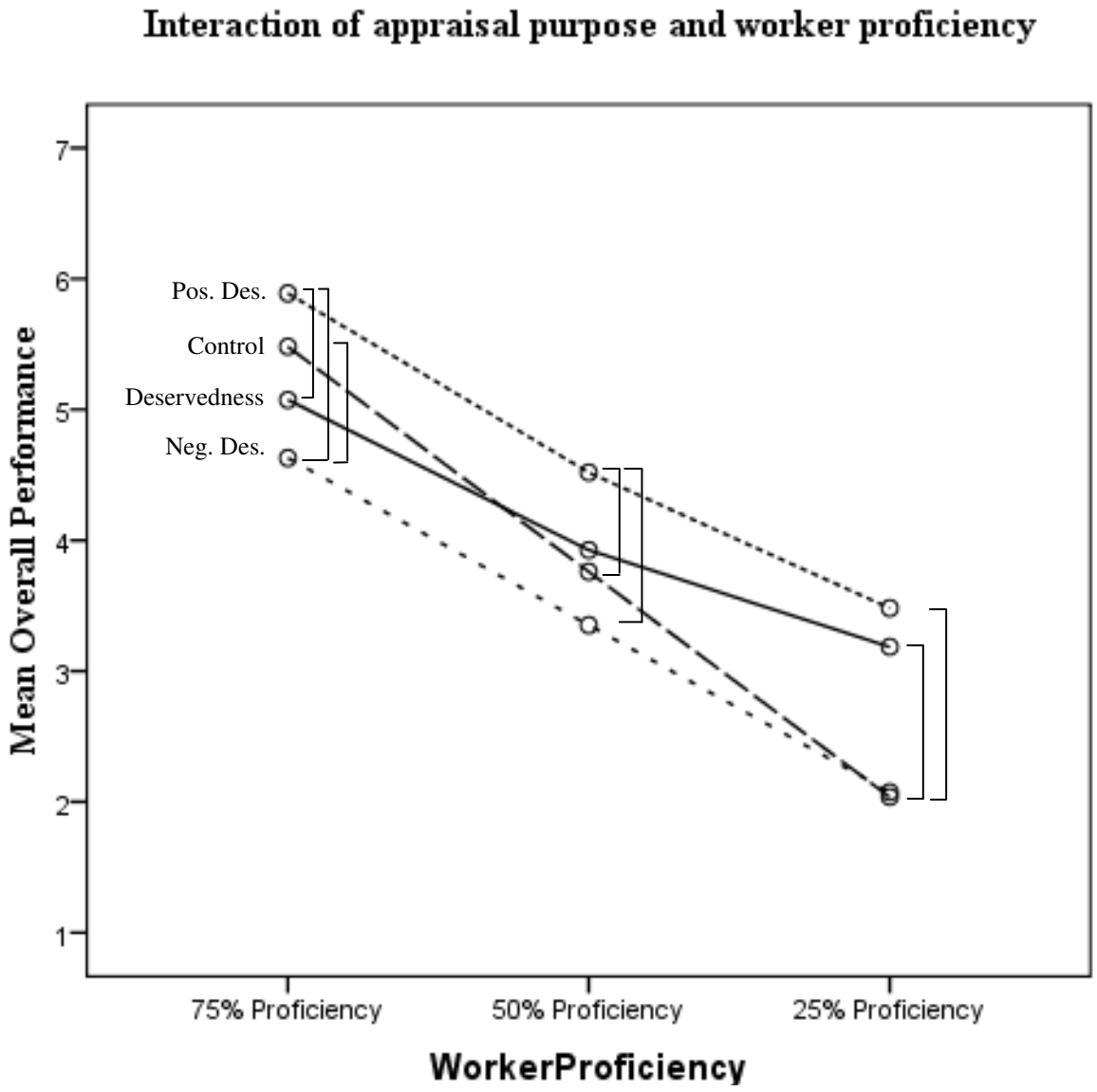

Figure 1: Appraisal purpose and worker proficiency

Note: All mean differences within appraisal condition were significantly different. Significant differences between appraisal conditions within each proficiency level are noted with black bars. Pos. Des. refers to Positive Designation, Neg. Des. refers to Negative Designation.

Figure 2: Appraisal purpose and expertise with strict definition at high proficiency

Note: Overall performance ratings for the 75% proficiency worker for each appraisal condition and expertise level (using the stricter definition of expertise). This interaction was not significant.

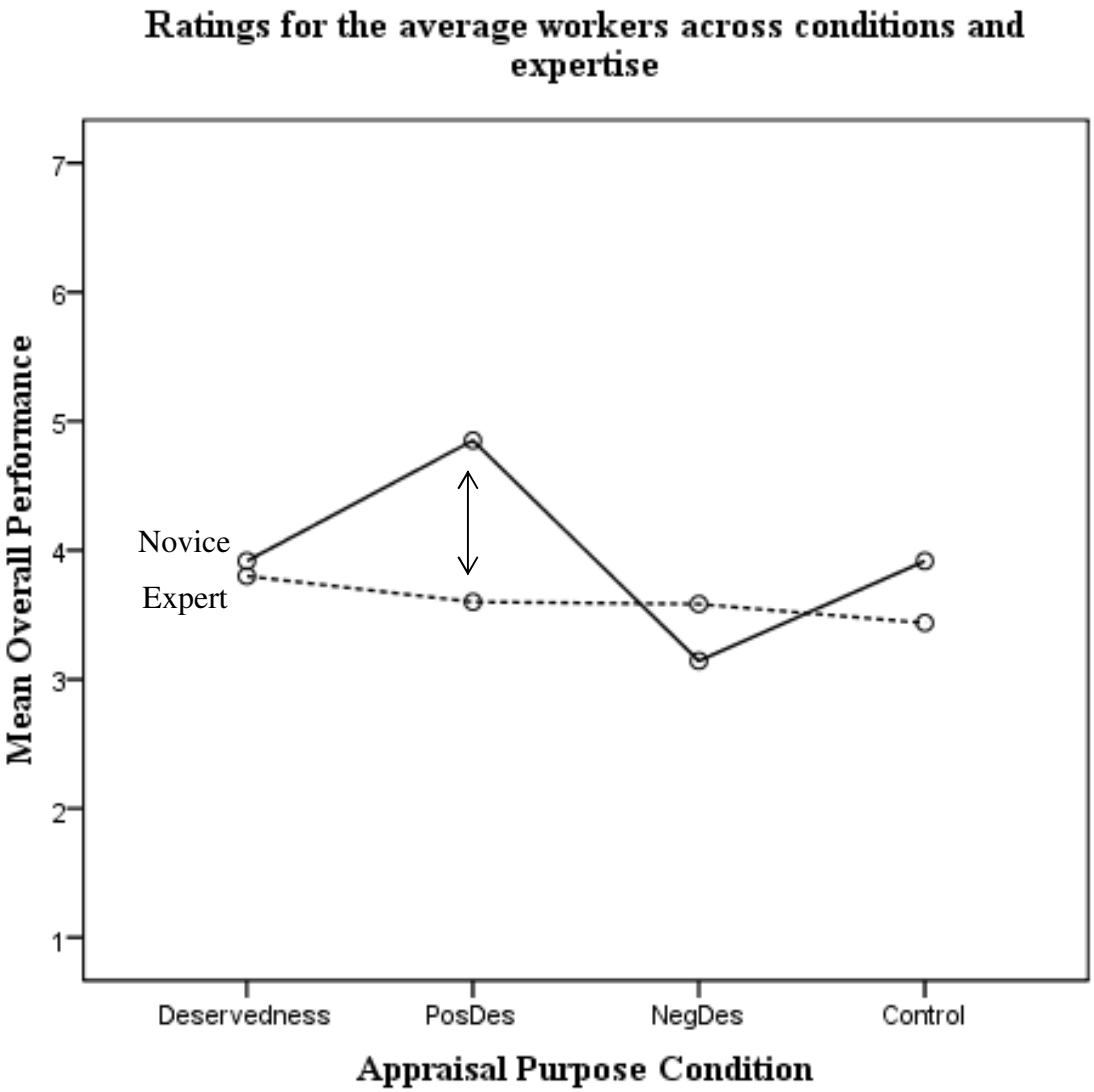

Figure 3: Appraisal purpose and expertise with strict definition at medium proficiency

Note: Overall performance ratings for the 50% proficiency workers for each appraisal condition and expertise level (using the stricter definition of expertise). This interaction was significant, $F(3, 45) = 2.93$, $p = .044$. Arrows indicate a significant difference, $p < .05$.

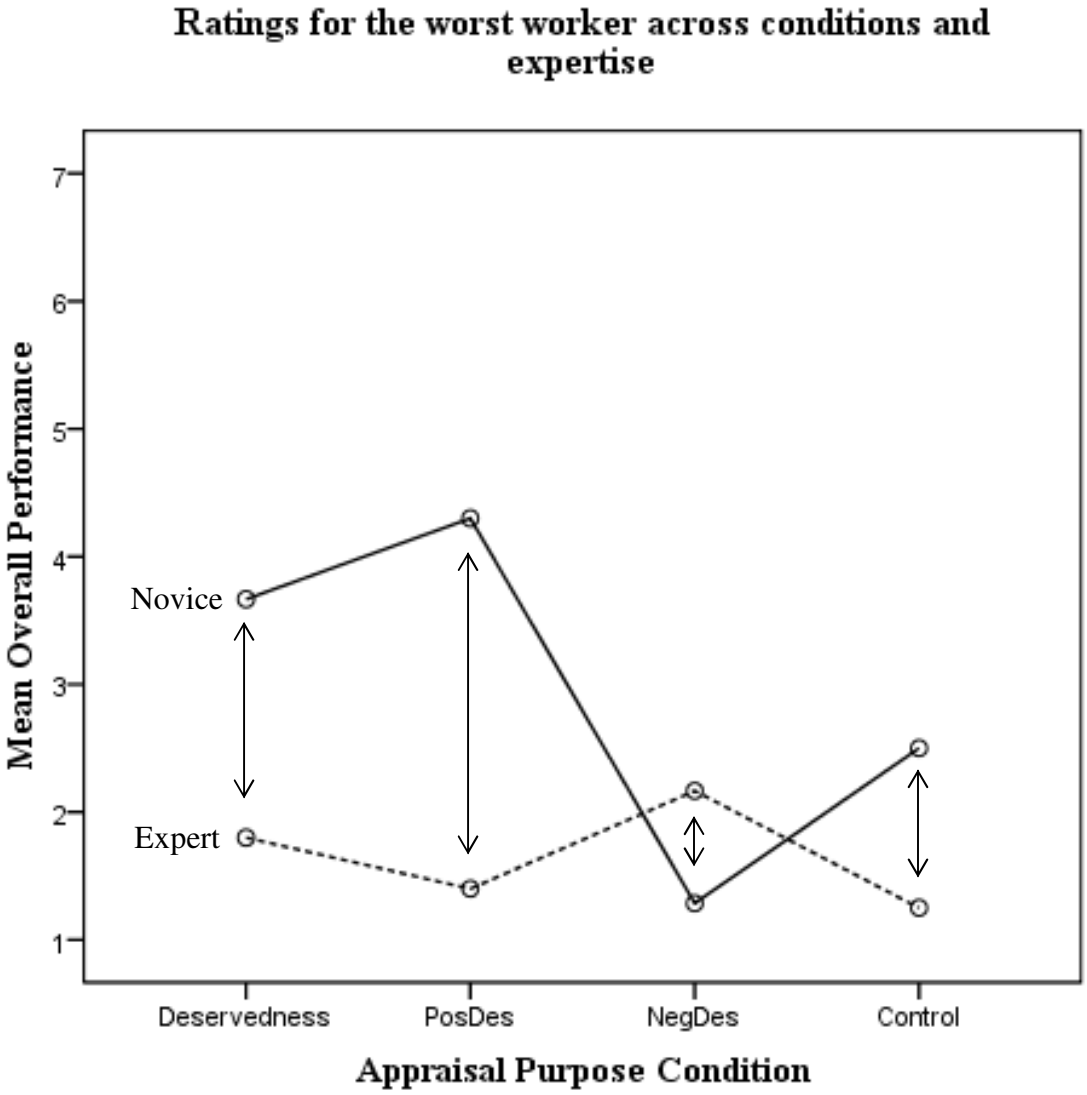

Figure 4: Appraisal purpose and expertise with strict definition at low proficiency

Note: Overall performance ratings for the 25% proficiency workers for each appraisal condition and expertise level (using the stricter definition of expertise). This interaction was significant, $F(3, 45) = 12.95$, $p < .001$. Arrows indicate a significant difference, $p < .05$.

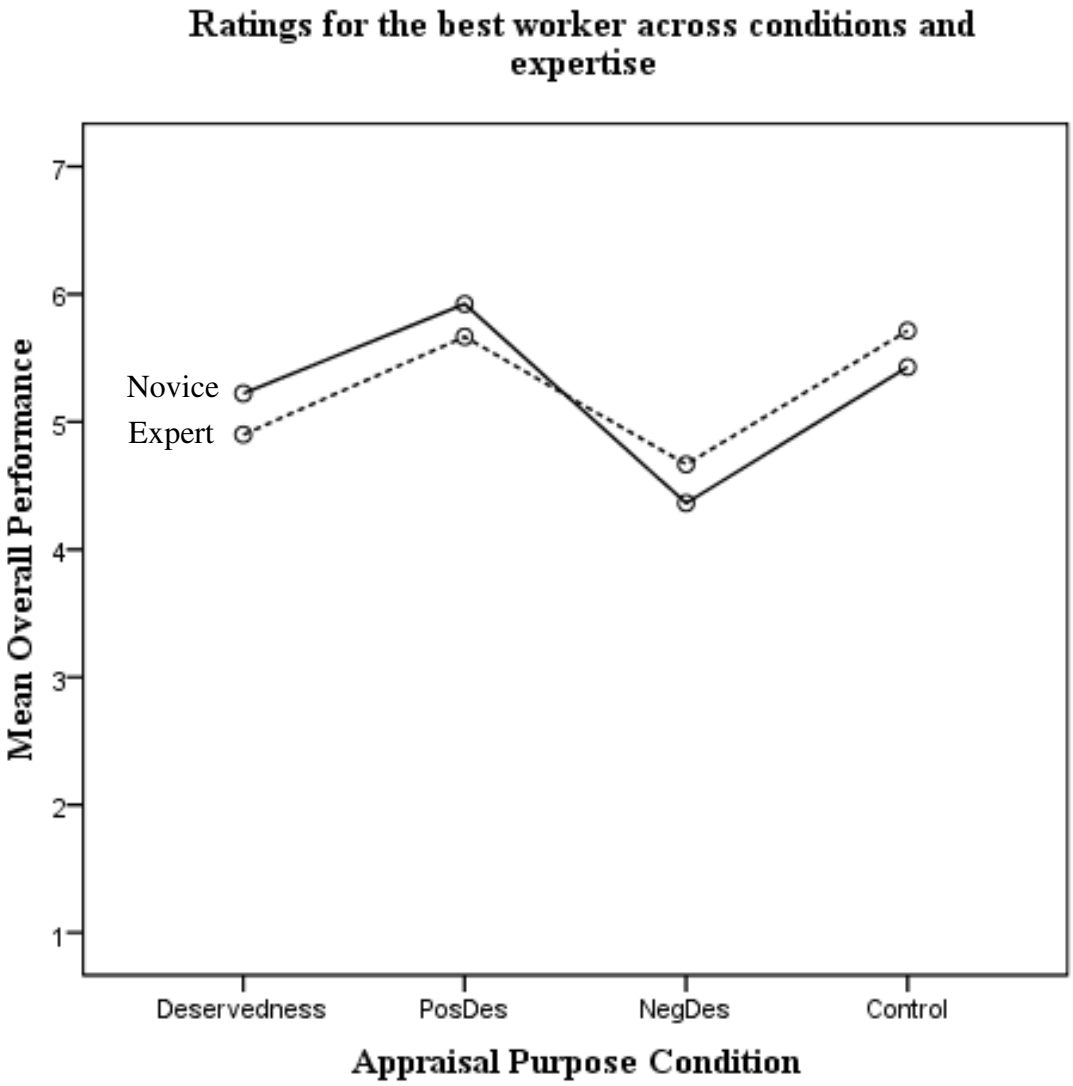

Figure 5: Appraisal purpose and expertise with loose definition at high proficiency

Note: Overall performance ratings for the 75% proficiency worker for each appraisal condition and expertise level (using the looser definition of expertise). This interaction was not significant.

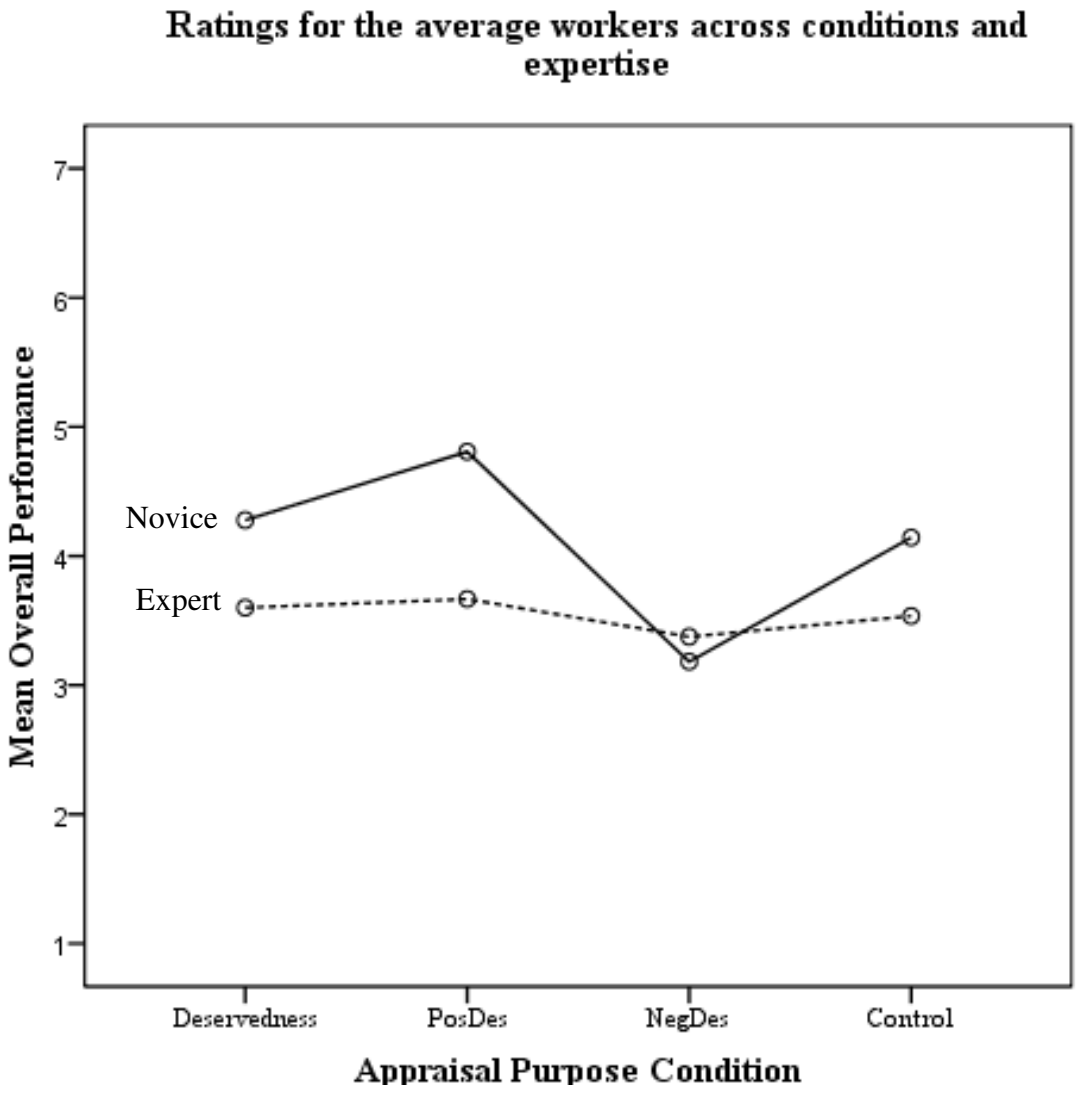

Figure 6: Appraisal purpose and expertise with loose definition at medium proficiency

Note: Overall performance ratings for the 50% proficiency workers for each appraisal condition and expertise level (using the looser definition of expertise). This interaction was not significant.

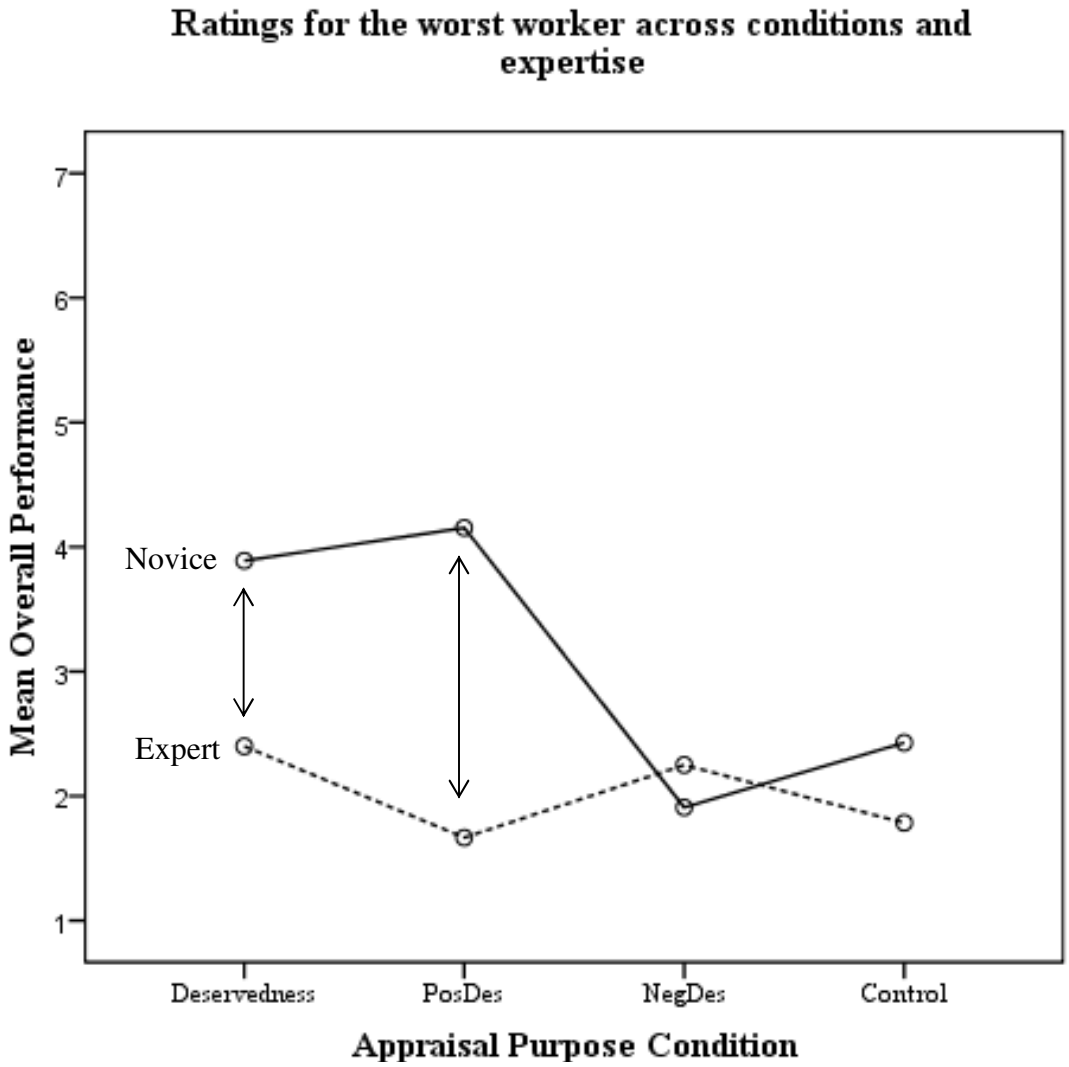

Figure 7: Appraisal purpose and expertise with loose definition at low proficiency

Note: Overall performance ratings for the 25% proficiency worker for each appraisal condition and expertise level (using the looser definition of expertise). This interaction was significant, $F(3,74) = 6.22$, $p = .001$.

Appendix A – Demographic Questionnaire (All Conditions)

Please answer the following questions as honestly and accurately as possible:

1. Sex: ☐ Male ☐ Female

2. Race: ☐ African-American/Black ☐ Native American ☐ White ☐ Asian ☐ Hispanic ☐ Other _____

3. Age: _____

4. Education Status: ☐ Freshman ☐ Sophomore ☐ Junior ☐ Senior ☐ Grad Student ☐ Not a student

5. Employment Status:
☐ Currently part-time (<40 hrs/week)
☐ Currently full-time (40+ hrs/week)
☐ Currently unemployed, but have been employed in the past
☐ Never been employed

6. If you are currently employed, either part-time or full-time, what is your job title?

_____

7. On the scale below, please <u>circle the number best representing your level of familiarity</u> with cardiopulmonary resuscitation (CPR).

<--|----------|----------|----------|----------|----------|----------|-->
(1) (2) (3) (4) (5) (6) (7)

1 = I have no experience with CPR and couldn't possibly perform it.
4 = I have some experience, but wouldn't feel very comfortable performing it
7 = I am very familiar with proper CPR technique and am fully prepared to perform it if necessary

7. Please indicate which of the following most accurately reflects the status of your certification in CPR/First Aid.

☐ Currently certified to perform and train others in CPR/First Aid
☐ Currently certified to perform CPR/First Aid
☐ Have been certified in the past, but am no longer current
☐ Have taken a CPR/First Aid class, but never been certified
☐ Have never taken a CPR/First Aid class of any kind

8. If you have ever been certified where did you get your certification? (e.g. Red Cross in Topeka)

_____

Appendix B – Written CPR Guidelines

How to evaluate CPR performance

Adult – Compressions

- 30 compressions
- Straight elbows
- Hell of hand pressed in middle of chest, between nipples
- Fingers interlaced
- Compression of 1 ½ to 2 inches
- Steady rate of compression
    - Similar to beat of "Stayin' Alive" by the Bee Gees

Adult – Rescue Breaths

- Pinching nose
- Head tilted back
- Chest should rise
- Fingers on chin, away from throat
- Two one-second breaths

Infant – Compressions

- 30 compressions
- Tips of two fingers
- Middle of chest, between nipples
- Fingers pointed straight down
- Compression of 1/3 to 1/2 the depth of the chest
- Steady rate of compression
    - Again… "Stayin' Alive"

Infant – Rescue Breaths

- Completely cover nose and mouth with your mouth
- Only blow enough to make chest rise

These guidelines are based on the American Heart Association's CPR Anytime Personal Learning Program.

Appendix C – Positive Designation Session 1 Rating Form

CODE: ____

Based on her performance, please indicate which person should be rewarded for their performance with a recommendation for admission into the NIH study by placing an X on the line next to her name.

___ Jennifer _____ Molly _____Katie _____Caitlin

Appendix D – Negative Designation Session 1 Rating Form

CODE: ____

Based on her performance, please indicate which person should be eliminated from the applicant pool for the NIH study due to poor performance by placing an X on the line next to her name.

___ Jennifer _____ Molly _____Katie _____Caitlin

Appendix E – Deservedness & Control Session 1 Rating Form

CODE: ____

Please rate each person on the following scale, with the anchors listed below, indicating whether you think her performance warrants her admission into the NIH study or if she should be eliminated from the applicant pool. Please place a whole number on the line indicating your rating on the line next to the person's name.

<--|----------|----------|----------|----------|----------|----------|-->
(1) (2) (3) (4) (5) (6) (7)

1 = Definitely should be eliminated from consideration for the NIH study
4 = Adequate performance, but not the best candidate
7 = Would be an ideal candidate for the NIH study

___ Jennifer _____ Molly _____Katie _____Caitlin

Appendix F –Session 2 Rating Form (All Conditions)

CODE _____

Jennifer **Molly** **Katie** **Caitlin**




Recall the performance videos you viewed two days ago. Using the following scale, with the anchors listed below, please rate each person's performance, both individually by task, and overall. You can consult the pictures at the top of this page to jog your memory about each person, but you will not have the opportunity to view their performance again.

<--|----------|----------|----------|----------|----------|----------|-->
(1) (2) (3) (4) (5) (6) (7)

1 = Very poor
4 = Average
7 = Outstanding

Name	**Jennifer**	**Molly**	**Katie**	**Caitlin**
Overall Performance	_____	_____	_____	_____
Adult Compressions	_____	_____	_____	_____
Adult Breaths	_____	_____	_____	_____
Infant Compressions	_____	_____	_____	_____
Infant Breaths	_____	_____	_____	_____

Appendix G – Positive Designation Session 1 Protocol

Have participants sign in and give them the Informed Consent form and appropriate coded response packet. Once they have looked over and signed the Informed Consent form, have them fill out the first page of the packet, the demographics questionnaire. Once they are done…

> Thank you for participating in this study. This project has multiple purposes. First, as part of selecting participants for a different study sponsored by a grant from the National Institutes of Health, we would like to obtain ratings from a diverse group of objective reviewers. Thanks to the grant, we are offering significant monetary compensation for participants in that study, which is why we are going through a process of carefully selecting the best applicants.
>
> The study requires participants who are capable in CPR. So, part of the application process is to perform CPR on an adult and infant dummy. You will view videos of each person doing CPR and evaluate them.
>
> A secondary purpose of this study is to conduct some research on the way people rate performance.
>
> So I'll be showing you a clip from a training video on CPR that will give you an idea of how to evaluate CPR performance. The same basic information is provided on the second page of your packet. It is important that you understand simply watching this clip will not make you certified to perform CPR yourself, but will hopefully provide you with enough information to make an informed rating.

Play CPR skills DVD.

> Now please look over the written guidelines to make sure you are familiar with what to look for.

Cue up CPR performance DVD.

> You will now view four of our applicants performing CPR on both an adult and infant dummy. Your job is to help us identify the best performer, who you think most deserves to be included in the NIH study. Their names will appear before they come on the screen, so make sure you pay attention during transitions. If there are no questions, let's begin.

Appendix H – Negative Designation Session 1 Protocol

Have participants sign in and give them the Informed Consent form and appropriate coded response packet. Once they have looked over and signed the Informed Consent form, have them fill out the first page of the packet, the demographics questionnaire. Once they are done…

> Thank you for participating in this study. This project has multiple purposes. First, as part of selecting participants for a different study sponsored by a grant from the National Institutes of Health, we would like to obtain ratings from a diverse group of objective reviewers. Thanks to the grant, we are offering significant monetary compensation for participants in that study, which is why we are going through a process of carefully selecting the best applicants.
>
> The study requires participants who are capable in CPR. So, part of the application process is to perform CPR on an adult and infant dummy. You will view videos of each person doing CPR and evaluate them.
>
> A secondary purpose of this study is to conduct some research on the way people rate performance.
>
> So I'll be showing you a clip from a training video on CPR that will give you an idea of how to evaluate CPR performance. The same basic information is provided on the second page of your packet. It is important that you understand simply watching this clip will not make you certified to perform CPR yourself, but will hopefully provide you with enough information to make an informed rating.

Play CPR skills DVD.

> Now please look over the written guidelines to make sure you are familiar with what to look for.

Cue up CPR performance DVD.

> You will now view four of our applicants performing CPR on both an adult and infant dummy. Your job is to help us identify the worst performer, who you think most deserves to be removed from the pool of applicants. Their names will appear before they come on the screen, so make sure you pay attention during transitions. If there are no questions, let's begin.

Appendix I – Deservedness Session 1 Protocol

Have participants sign in and give them the Informed Consent form and appropriate coded response packet. Once they have looked over and signed the Informed Consent form, have them fill out the first page of the packet, the demographics questionnaire. Once they are done…

> Thank you for participating in this study. This project has multiple purposes. First, as part of selecting participants for a different study sponsored by a grant from the National Institutes of Health, we would like to obtain ratings from a diverse group of objective reviewers. Thanks to the grant, we are offering significant monetary compensation for participants in that study, which is why we are going through a process of carefully selecting the best applicants.
>
> The study requires participants who are capable in CPR. So, part of the application process is to perform CPR on an adult and infant dummy. You will view videos of each person doing CPR and evaluate them.
>
> A secondary purpose of this study is to conduct some research on the way people rate performance.
>
> So I'll be showing you a clip from a training video on CPR that will give you an idea of how to evaluate CPR performance. The same basic information is provided on the second page of your packet. It is important that you understand simply watching this clip will not make you certified to perform CPR yourself, but will hopefully provide you with enough information to make an informed rating.

Play CPR skills DVD.

> Now please look over the written guidelines to make sure you are familiar with what to look for.

Cue up CPR performance DVD.

> You will now view four of our applicants performing CPR on both an adult and infant dummy. Your job is to rate each applicant individually on whether you think they should be included in the NIH study or eliminated from further consideration. Their names will appear before they come on the screen, so make sure you pay attention during transitions. If there are no questions, let's begin.

Appendix J – Control Session 1 Protocol

Have participants sign in and give them the Informed Consent form and appropriate coded response packet. Once they have looked over and signed the Informed Consent form, have them fill out the first page of the packet, the demographics questionnaire. Once they are done…

> Thank you for participating in this study. This project has multiple purposes. First, as part of selecting participants for a different study sponsored by a grant from the National Institutes of Health, we would like to obtain ratings from a diverse group of objective reviewers. Thanks to the grant, we are offering significant monetary compensation for participants in that study, which is why we are going through a process of carefully selecting the best applicants.
>
> The study requires participants who are capable in CPR. So, part of the application process is to perform CPR on an adult and infant dummy. You will view videos of each person doing CPR and evaluate them.
>
> A secondary purpose of this study is to conduct some research on the way people rate performance.
>
> So I'll be showing you a clip from a training video on CPR that will give you an idea of how to evaluate CPR performance. The same basic information is provided on the second page of your packet. It is important that you understand simply watching this clip will not make you certified to perform CPR yourself, but will hopefully provide you with enough information to make an informed rating.

Play CPR skills DVD.

> Now please look over the written guidelines to make sure you are familiar with what to look for.

Cue up CPR performance DVD.

> You will now view four of our applicants performing CPR on both an adult and infant dummy. Afterward, you will evaluate them. Their names will appear before they come on the screen, so make sure you pay attention during transitions. If there are no questions, let's begin.