THESIS

RATING ACCURACY AND COGNITIVE LOAD ASSOCIATED WITH THE DISTRIBUTIONAL ASSESSMENT MODEL

Submitted by

Adam Vanhove

Department of Psychology

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Spring 2011

Master's Committee:

Advisor: Alyssa Mitchell Gibbons

Benjamin A. Clegg
Chris A. Henle
George C. Thornton III

ABSTRACT

RATING ACCURACY AND COGNITIVE LOAD ASSOCIATED WITH THE DISTRIBUTIONAL ASSESSMENT MODEL

This study examined both the interrater agreement and true score accuracy associated two different types of response formats, one using the traditional assessment (TA) approach, and the other using the distributional assessment (DA) approach. In addition, proponents of the DA response format have proposed that DA users experience less cognitive load than TA users (e.g., Kane, 2000), however, this has not been empirically examined until now. Findings suggest 1) greater interrater agreement among DA users, 2) higher true score accuracy for DA users despite minimal practical significance, and 3) DA users actually experienced significantly more cognitive load than TA users. Finally, a mediational hypothesis was tested to examine whether response format led to experienced cognitive load which, in turn, led to differences in true score accuracy. No evidence was found for this mediational hypothesis.

TABLE OF CONTENTS

INTRODUCTION

Rating Accuracy and Cognitive Load Associated with the Distributional Assessment Model

Human resource (HR) decisions are vital to organizational success (Feldman, 1981; Judge & Ferris, 1993; Thornton & Rupp, 2005). The most commonly used internal HR decision making tool is performance ratings (e.g., Borman, 1975; Guion, 1965; Landy & Rastegary, 1988) – that is, performance indicators requiring human judgment. A performance rating may be as simple as a restaurant patron using a customer service card, but even this simple data can drive organizational decisions. For example, if John, a server, receives positive ratings from his customers, the manager may decide to schedule him more often or during peak hours. If John continues to receive positive customer ratings he may eventually be promoted to shift manager. Many performance ratings are more complex; for example, completing an annual performance appraisal form for a subordinate. Here, the rater may be asked to provide ratings of multiple performance dimensions (e.g., teamwork, customer service, leadership), and these may be used to evaluate the subordinate's performance across several months. Again, such ratings have consequences for organizations and employees. For example, if Terry, a customer service representative, is rated as working poorly with others and having poor relations with customers, she may be reprimanded, transferred to a position in the company requiring little interpersonal interaction, or even let go if problems persist.

Despite the popularity of performance ratings, their use comes at a cost – susceptibility to human judgment error (e.g., Murphy & Cleveland, 1995). For example, John, the server, may have mixed up customers' drink orders and given them the wrong change at the end of the night, but this may have been overlooked because John was polite, handsome, and funny. Similarly, Terry, the customer service rep, may have actually interacted quite well with customers and coworkers across the appraisal period with the exception of one or two recent interactions. However, because these were fresh in the supervisor's mind, the supervisor may have weighed these interactions more heavily, and, as result, rated Terry's performance inaccurately.

Concerns about rater error in performance ratings have existed in research and practice for over a century (see Austin & Villanova, 1992). Although rater error will likely not be eliminated, research aimed at reducing it has been met with at least some success. Strategies that appear to reduce rater error include providing rater training (Bernardin & Buckley, 1981; Woehr & Huffcutt, 1994), increasing the experience level of raters (e.g., Kolk, Born, van der Flier, & Olman, 2002), and reducing the number of dimensions on which individuals are evaluated (Gaugler & Thornton, 1989).

One popular strategy for reducing rater error is to change the format of the ratings (e.g., behavioral checklists [Hennessy, Mabey, & Warr, 1998; Reilly, Henry, & Smither, 1990]; BARS [Smith & Kendall, 1963]). The present study aims to examine one particularly innovative type of rating format: distributional assessment (DA; Kane, 1986, 2000; Kane & Woehr, 2006). Previous research on DA suggests that its use may result in less rater error and greater rater accuracy (e.g., Deadrick & Gardner, 1997; Kane, 2000; Woehr & Miller, 1997). Distribution assessment proponents have theorized that the

increased rater accuracy associated with DA is the result on a reduced level of cognitive load being placed on the rater (e.g., Edwards & Woehr, 2007; Kane, 2000). However, this assumption has yet to be directly tested. Thus, this study aims to add to our understanding of why DA may lead to more accurate ratings by examining the role of cognitive load experienced by raters and explicitly tests whether cognitive load mediates the relationship between response format and rater accuracy.

The Distributional Assessment (DA) Model

To understand the potential value of DA it is first important to understand how performance ratings are traditionally reported. As the name suggests, the traditional rating response format (TA) is the most commonly used in practice. Traditional assessment requires that users report a mean rating to summarize the rating target's performance using a Likert-type scale. In the case of John, the server, this would likely involve a single performance dimension – customer service. Here, the dinner patrons would be required to mark the rating category (e.g., 1 = very poor, to 5 = very good) that best represents John's overall performance during their experience at the restaurant. A simplified example of the TA response format is shown in Figure 1 with a mean rating of "2".

Figure 1. Example of a TA Response Format.

Instructions: Please rate the employee's service.

1-Very poor	2-Poor	3-Acceptable	4-Good	5-Very good
_____	__X__	_____	_____	_____

However, research suggests that performance varies across time (e.g., Beal, Weiss, Barros, & McDermid, 2005; Reb & Cropanzano, 2007). This was the driving theory behind Kane's (1986) development of the DA format. Indeed, Kane suggests that performance is iterative – that is, that many of the behaviors relevant to performance are produced on multiple occasions during the rating period – creating a distribution of performance. Moreover, Kane (1986, 2000) and Kane and Woehr (2006) suggest that the mean ratings associated with TA do not lend themselves well to capturing the iterative nature of performance.

Recall the example of Terry. Terry's performance appraisal is intended to reflect her performance across a six month period. It is unlikely that Terry performed at exactly the same level every day for six months. In other words, it is highly unlikely that Terry simply continuously met her supervisor's expectations. Instead, she likely exceeded them at times, met them at times, and at times failed to meet them. This suggests that Terry's performance was distributed over the different performance levels across time. Such a distribution of performance may be difficult to capture simply through mean ratings. Through the use of DA, the variability in Terry's performance may be better captured when the supervisor is allowed to report the distribution of Terry's performance across all levels on the rating scale. Figure 2 depicts just that. Here, the supervisor estimates the frequency of Terry's performance for each category on the scale.

Figure 2. Example of DA Response Format.

1-Very poor	2-Poor	3-Acceptable	4-Good	5-Very good
__15_%	__15_%	__30_ %	__30_ %	__10_%

On the surface, the DA rating format appears similar to the traditional format, with categories representing different levels of performance. However, Kane (1986, 2000) suggests that DA differs from TA in one very important way – instead of requiring the rater to mentally average the range of target behaviors into a single mean rating as is the case when using TA, DA asks raters to report the frequencies with which targets perform at each level on the rating scale. This is often done by asking raters to note the percentage of time the rating target performed at each level on the rating scale.

In Figure 2, Terry's supervisor documented that Terry performed very poorly 15 percent of the time within the rating period, very well 10 percent of the time, and so on. Mean ratings can be easily calculated by averaging frequencies across the different levels (the distribution in Figure 2 results in a mean rating of *3.05*), so all of the information provided in TA is available in DA as well. However, DA provides additional information not available through the use of TA regarding the distribution of performance (Kane, 1986).

This additional information regarding the frequency and distribution of performance across the rating scale may be very valuable to HR decision makers, and may be sufficient reason in itself to prefer DA to TA (Kane, 1986, 2000; Kane & Woehr, 2006). However, some research also suggests that DA leads to greater rating accuracy when compared to TA. In the next section, I review this research and then consider its limitations.

Research on DA and rating accuracy. Early laboratory-based research provided only mixed support for DA increasing rater accuracy. The first study to

compare rating accuracy between TA and DA was a laboratory experiment conducted by Jako and Murphy (1990). Findings from that study suggested that DA provided no advantage over TA in interrater agreement. Steiner et al. (1993) also examined interrater agreement and found no differences in participant agreement between TA and DA rating conditions. More recently, Woehr and Miller (1997) found significantly better goodness-of-fit among DA user ratings than TA user ratings, and concluded that the lower magnitude of measurement error associated with the DA model represented less rating error.

Research comparing the TA and DA rating formats also exists using field samples. Deadrick and Gardner (1997) examined the relationship of DA and TA ratings to objective performance measures in a sample of sewing machine factory supervisors. They found that mean performance ratings were similar, but that when rating variable performance, DA users' ratings were more closely related to the objective performance criteria than were the TA ratings. The only other study in an applied setting was conducted by Fox, Bizman, and Garti (2005). Using a sample of computer programmers, they also found DA and TA mean ratings to be similar (r=.81). However, the interrater reliability of stereotype accuracy and differential accuracy (two components of accuracy developed by Cronbach [1955]) were higher in DA ratings than in TA ratings.

These findings suggest that DA has considerable potential to improve the rating process. However, they must be viewed with some caution. All of the studies described above (with the exception of Deadrick and Gardner [1986]) have conceptualized rater accuracy through different measures of interrater agreement (e.g., Fox et al., 2005; Jako & Murphy, 1990; Steiner et al., 1993). Cronbach's (1955) four components of accuracy,

as used by Fox et al., offer the most comprehensive picture of rater agreement.. These components include: *elevation*, or the grand mean interrater agreement, *differential elevation*, or interrater agreement for each rating target, *stereotype accuracy*, or interrater agreement for each skill dimension, and *differential accuracy*, or the interrater agreement for each rating target x skill dimension combination. For each of the above components each participant's ratings are compared to the rest of the participants – there is no absolute or true score involved..

Interrater agreement measures can provide some important information that may suggest that one response format leads to more accurate ratings than another, however, interrater agreement is not a true indicator of rater accuracy. Thus, I included expert panel-developed true score differences to directly measure rater accuracy. No study has examined DA and TA using true scores as rating accuracy criteria. Thus it remains unclear whether or not using DA actually leads to greater rater accuracy as opposed to TA.

Separate hypotheses were developed to assess different aspects of rater accuracy. The decision to examine these two sets of dependent variables was based on the inherent differences in the dependent variables. On the one hand, Cronbach's components of accuracy essentially reflect four different estimates of interrater agreement (e.g., across rating dimensions, across rating targets). On the other hand, true score accuracy reflects the difference between participants' ratings of each rating target and the corresponding true score developed by the expert rating panel. Thus, the first hypothesis presented below aims to establish whether or not participants using the DA response format have higher agreement in their ratings than those using the TA response format. Although not

a true measure of accuracy, a comparison of interrater agreement using Cronbach's components of accuracy as measures of interrater agreement are used in order to compare findings of this study to previous DA research. The second hypothesis aims to directly test whether or not DA users are more accurate than TA users in the context of the current study. Thus, the first set of hypotheses is as follows:

H1a: Ratings made using DA, in comparison to TA, will result in higher interrater agreement in terms of stereotype accuracy and differential accuracy[1].

H1b: Ratings made using DA, in comparison to TA, will result in lower true score differences.

The advantage of DA over TA appears to be stronger when the rating target's performance is inconsistent (e.g., Deadrick & Gardner, 1997; Steiner et al., 1993). Indeed, Deadrick and Gardner (1997) failed to find significant differences between DA and TA ratings under conditions of consistent performance. However, they did find DA ratings to be more accurate than TA ratings under conditions of inconsistent performance. These findings are not surprising. Returning to the example of Terry, Terry's supervisor would likely not have difficulty rating Terry's performance accurately using TA if Terry's performance did not vary across the appraisal period. In other words, there would be less potential for Terry's supervisor to let, for example, Terry's most recent or initial performance to erroneously influence the supervisor's rating, because all performance would be at the same level. However, if Terry's performance was

[1] Interrater agreement in this study is reflected by Cronbach's four components of accuracy. The specific components denotes in H1a are line with the findings from Fox et al. (2005).

inconsistent, the potential for isolated, uncharacteristic behaviors to erroneously influence Terry's supervisor's ratings may be increased. In line with past findings regarding performance inconsistency and DA rating accuracy it is expected that the use of DA will be especially advantageous in situations when target performance is inconsistent:

H2a: The difference in interrater agreement between TA and DA will be more pronounced as target performance variability increases.

H2b: The difference in true score accuracy between TA and DA will be more pronounced as target performance variability increases.

Distributional Assessment and Cognitive Load

DA proponents have argued that DA increases rater accuracy because it results in a reduction in cognitive load experienced during the rating task (Kane, 2000; Kane & Woehr, 2006). Cognitive load can be described as the amount of cognitive resources that an individual commits to information processing at any given point[2] (Paas & van Merrienboer, 1994). Individuals' information processing capacity is limited, and one's cognitive load represents the proportion of this capacity that is utilized (Sweller, 2006). In the context of performance ratings, cognitive load is experienced by raters throughout

[2] It should be noted that more recently cognitive psychological theory has referred to cognitive load alternatively as activated long-term memory (see Ruchkin, Grafman, Cameron, & Berndt, 2003 for this alternative view), an approach which includes a more complex view of working memory. Although theory behind these two approaches differs, it is important to highlight that the key idea of both approaches is the same: that individuals' information processing capacity is limited. For simplicity, the term cognitive load is used throughout the remainder of this paper to refer to the amount of mental capacity being used at a given time.

the rating process to the extent to which raters must consciously process information regarding target performance.

Cognitive load and rater error. There appears to be a general consensus among scholars that rater error is a function of both the complexity of the rating task and the insufficient information processing capacity of raters (e.g., Bycio, Alvares, & Hahn, 1987; Murphy & Cleveland, 1995; Reilly et al., 1990; Schleicher et al., 2002; Zedeck, 1986). Researchers also generally agree that the performance rating process is complex and have accepted that it may be unrealistic to expect raters to complete this task accurately and reliably (Bycio et al., 1987; Reilly et al., 1990). In other words, it is suggested that rating performance places a high level of cognitive load on raters (see Murphy & Cleveland, 1995; Zedeck, 1986).

As mentioned at the outset of this paper, a wealth of research has examined characteristics of the performance rating task as determinants of rater accuracy (e.g., Kolk et al., 2002; Reilly et al., 1990). Interestingly, much of this literature has focused on simplifying the rating task and reducing the cognitive load experienced by raters. For example, Gaugler and Thornton (1989) proposed that reducing the number of rating dimensions (e.g., oral communication, teamwork) would reduce the cognitive complexity of the rating task. A second example is highlighted in the rater training research (e.g., Bernardin & Buckley, 1981; Cardy & Keefe, 1994; Day & Sulsky, 1995; Pulakos, 1984; Schleicher et al., 2002; Woehr & Huffcutt, 1994). This research implies that, because rater training better prepares raters for the rating task, they experience less cognitive load in the process (e.g., DeNisi & Peters, 1996; Pulakos, 1988). Given the abundance of the research that has been conducted under the assumption that manipulating the rating task

reduces raters' cognitive load, it is surprising that none of this research has directly examined cognitive load.

Cognitive load and DA. The idea of cognitive load as a key determinant of rater accuracy is particularly important in theory surrounding DA. Research examining the use of event frequency-based responding (the basis on which DA was developed) in the cognitive and evolutionary psychology domains suggests that a frequency-based response format (like DA) should reduce cognitive load. This body of research has led DA proponents to speculate that DA reduces raters' cognitive load during the rating process in both quantitative and qualitative ways (e.g., Edwards & Woehr, 2007; Kane, 2000; Kane & Woehr, 2006). Quantitatively, raters' cognitive load is thought to be reduced through DA's reliance on event frequencies. Cognitive research has suggested that the frequencies of events are automatically encoded and recalled (see Hasher & Zacks, 1984) requiring few cognitive resources. Further support can be inferred from the evolutionary psychology literature which suggests that the automatic and unconscious encoding of events is an adaptive trait of the human mind (Cosmides & Tooby, 1996). As these processes are believed to be automatic, it becomes plausible that raters using DA experience quantitatively less cognitive load than raters who must expend cognitive resources to, at least some degree, actively encode and retrieve target behaviors.

Qualitatively, DA differs from TA in that DA does not require raters to aggregate behavioral examples into a single mean rating, and this too may reduce the cognitive load placed on raters. Recall again the example of Terry, and her supervisor having to mentally aggregate the entire range of Terry's customer service-relevant behaviors into a single mean rating when using TA. However, when using DA, the aggregation stage is

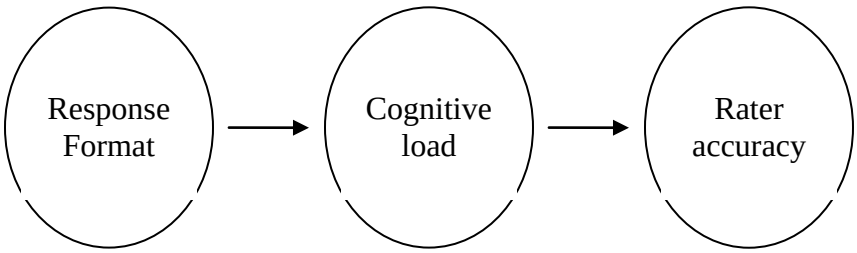
largely bypassed as raters simply report the automatically encoded frequencies of performance for each rating category (Kane, 2000). Findings in the cognitive literature support this. For example, Sedlmeier, Hertwig, and Gigerenzer (1998) found that participants using frequency-based judgments expended fewer mental resources during aggregation .

Considering that DA is rooted in event frequency responding, it seems plausible to generalize these findings to DA in the context of performance ratings. However, it bears repeating that no performance rating research has measured cognitive load directly. As mentioned above, past rater accuracy research has traditionally implemented some variation to the rating task (e.g., rater note-taking) and assumed that findings of increased rater accuracy were the result of the simplification of the rating task and, in turn, a reduced level of cognitive load experienced by the rater. However, this may not be the case. In addition, cognitive load has not been previously examined in the DA literature. At this point, it is unclear whether or not the relatively brief rating tasks used in event frequency research differ from the complex judgments required when rating performance, especially over a long period of time (see Hastie & Park, 1986; Murphy, Balzer, Lockhart, & Eisenman, 1985; Murphy, Gannett, Herr, & Chen, 1986). Taken together, there appears to be a clear need to understand the effect that manipulating the rating task has on raters' experienced cognitive load. This study aimed to examine the effect of using DA on raters' cognitive load. Although there is little research to form a basis for the relationship between DA and cognitive load, that which does exists suggests that DA should reduce the amount of cognitive load experienced by raters. Based on this the second hypothesis is as follows:

H3: DA raters will experience less cognitive load than TA raters.

In the sections above, it was highlighted that a common thread in much of the research aimed at reducing rater error has focused on simplifying the task. Stated differently, this research has sought to improve accuracy by reducing raters' cognitive load. It was also highlighted that cognitive load has not been directly tested in this research. Thus, the implied model suggesting that cognitive load mediates the relationship between a given rating characteristic (e.g., the rating response format) and rater accuracy has only been implied up to this point. Although this mediational model has not been tested directly, much of this research has been successful in increasing rater accuracy (e.g., Gaugler & Thornton, 1989; Kolk et al., 2002), offering some indirect support for cognitive load having a mediating role. Nonetheless, the present study is the first to test the general mediational model (Figure 3) with regard to the use of DA.

Figure 3. Proposed Mediational Model.



In light of the arguments that DA both reduces rater cognitive load and increases rater accuracy (e.g., Kane, 2000), DA seems to be an appropriate choice to test this mediational model. Thus, the final hypothesis is as follows:

H4: Cognitive load will mediate the relationship between the response format (i.e., DA, TA) used and rater accuracy.

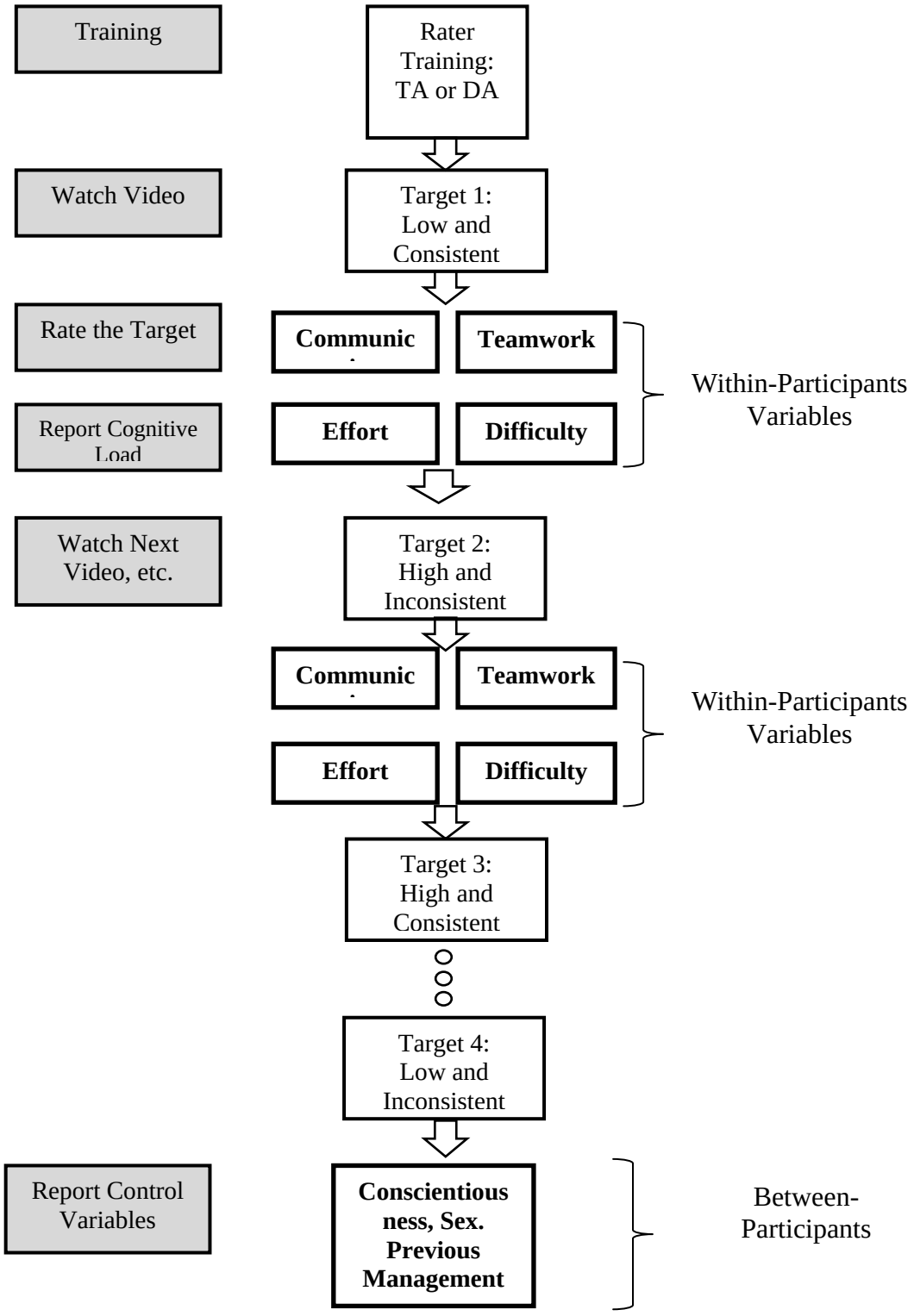
Current Study

This study examined participant raters' interrater agreement, true score accuracy, and experienced cognitive load.in rating simulation exercises. I used a between-participant design in which raters in a lab setting used either the TA or DA response format to evaluate different rating targets in a series of video-recorded simulation exercises. After viewing each video participants rated the target on two different skill dimensions – communication and teamwork. After rating each target, participants responded to two separate measures of cognitive load – cognitive effort and cognitive difficulty. Thus, raters rated four different rating targets and responded to the two measures of cognitive load four separate times. As a final step, participants were asked to complete a conscientiousness measure, and report their sex and whether or not they have had any previous experience in formally evaluating others. These final three measures were included as controls in order to rule out these variables as alternative explanations for potential group mean differences between the TA and DA ratings. This entire process is depicted in Figure 4.

Rater training was provided to all participants before the rating task began. The focus of the training was on familiarizing participants with the skill dimensions on which they were to rate targets, the behaviors which defined each level of performance, and the response format they were to use. Past research on DA has not examined DA's effect on rater accuracy when trained raters are used. Thus, although a student sample was used here, the rater training was intended to increase the realism of the study.

Following best practices for DA research as prescribed by Kane (2000), behaviorally anchored rating scales (BARS) were used rather than simple adjectival anchors (e.g., "good," "poor"). The goal of BARS is to use specific behavioral examples to define each rating category for a given skill dimension (Murphy & Cleveland, 1995). An abundance of research exists on BARS, but this research has not shown BARS to have a clear advantage over other scales (e.g., Borman, 1978; Murphy & Pardaffy, 1989). That said, many researchers suggest that using BARS in combination with other rating characteristics (e.g., rater training) may increase rating accuracy (Martinko & Gardner, 1985).

Figure 4. Schematic Overview of Study Design



METHOD

Participants

Two hundred five undergraduates from a large university in the Rocky Mountain region took part in this study. Participants took part in the study for partial fulfillment of course credit. Four participants were dropped from analysis due to having incomplete performance rating data. Thus, 201 participants (M=57, F=144) made up the final sample. Of the 201 participants, 33 identified themselves as having prior formal performance rating experience.

Sample size was based on the estimated effect sizes for DA and cognitive load for mediational tests using SEM. Past research on DA (e.g., Kane, 2000) suggests a small effect of rating format on accuracy (approximately .20). Due to the lack of research examining cognitive load and rating accuracy directly, the size of this effect is unclear. As a result, a conservative estimate of a small to medium effect size (d = .14-.26) for the cognitive load-to-accuracy path was used to estimate power. Using these two estimates, Fritz & McKinnon (2007) suggest a sample size of approximately 200.

Procedure

Each data collection session lasted approximately 90 minutes. At the outset of each session the researcher greeted the group and provided a general overview of the study. This was administered to the group as a whole and included instructions for participation.

During this portion of the data collection session, signed informed consent was obtained from all participants. Individuals were made aware that participation was completely voluntary and told they were free to withdraw at any time during the study. Participants were also informed that a monetary incentive would be awarded to the two most accurate ratings in both the TA and DA rating conditions. This incentive was offered to improve participant motivation to rate accurately. The remainder of the study was completed independently by the participants using laptop computers with head phones for audio.

Training. Before viewing targets in the videotaped simulation exercises and providing performance ratings, each participant took part in rater training. The training was approximately 20 minutes in length and was delivered through an automatized, interactive PowerPoint presentation. Questions regarding training content were presented throughout the PowerPoint to ensure that the participants understood the information.

This training covered two general content areas. The first portion of the training was adapted from the training materials for an AC already in use at the university. Here, participants were introduced to the AC method, the specific simulation exercise they would be viewing, and the skill dimensions on which they were to evaluate targets. The second section of the training aimed to familiarize the participants with the rating format (either DA or TA) that they would be using to rate the performance of targets. The training received by those in the TA and DA conditions was identical, with the exception of the single slide that described how to use the response format for rating target performance.

Rating performance. Upon completing the interactive training presentation, participants were asked to view four different videos of simulation exercises. The duration of each video was approximately eight minutes, resulting in a total viewing time of approximately 32 minutes. All participants viewed the same four simulation exercises and only the response format (i.e., DA and TA) and ordering of the videos differed between participants. Participants were directed to rate only one target during each exercise. Handouts were provided to all participants making clear which individual was the rating target in each video. Immediately after viewing each video, participants were asked to rate the target's performance on two different skill dimensions: teamwork and communication. The order in which participants viewed rating targets was one of two possible orders (target 1, 2, 3, 4 or target 4, 3, 2, 1). In addition, the order in which participants rated the two skill dimensions also differed (communication, then teamwork or teamwork, then communication).

After rating each target, participants then responded to two questions regarding the level of cognitive load they experienced while rating the target. After rating all four targets and responding to all cognitive load questions, participants were asked to respond to a conscientiousness measure, and indicate their sex and whether or not they had any previous experience in formal performance appraisal.

Materials

The simulation exercise and videos. Participants viewed the same simulation exercise in all four videos, with four different pairs of actors as the "participants" in the exercises. This was done in order to make more direct comparisons between ratings of targets in a single exercise rather than across largely different exercises. The simulation

exercise used in the current study was adapted from an existing AC. The exercise requires those being evaluated to plan a social event for incoming university freshmen.

Undergraduate and graduate students played the roles of participants in the simulation exercise. Each video included two individuals taking part in the exercise, and the target individual in each video was clearly identified through the handout. The target individual in each exercise was asked to convey behavior at one of two general levels of performance (i.e., high or low), and one of two levels of consistency (i.e., consistent or inconsistent). To aid the target in accurately portraying these combinations of performance characteristics, a list of behaviors to convey during the exercise was provided to targets in the four videotaped exercises (see Appendix D for example).

Traditional and distributional assessment response formats. The response scales for the current study were adapted from the BARS currently being used by raters for the university's AC. However, the original 7-point BARS was consolidated into a 4-point scale for both the DA and TA versions of the response scale. Excerpts of these scales are available in Appendices B and C, respectively. Participants were asked to rate each target on two dimensions: communication and teamwork.

Measures

Cognitive load. Cognitive load was measured using two separate single-item measures: "what was your level of mental effort during the task you just completed?" (Paas, 1992) and "how difficult was the task you just completed?" (Kalyuga, Chandler, & Sweller, 1999; Mayer & Chandler, 2001). These measures are popular in research on cognitive load and have been shown to be highly reliable (see Ayres, 2006). Participants

responded using a 9-point scale with "1" = low cognitive load and "9" = high cognitive load.

Conscientiousness. The nine conscientiousness items from the Big Five Inventory (BFI) (John, Donahue, & Kentle, 1991) were used to control for participants' conscientiousness. Participants responded to all nine items using a 5-point scale by marking how much they agreed with each statement ("1" = strongly disagree, "5" = strongly agree). The nine items used to measure conscientiousness are included in Appendix D. Conscientiousness scores were creating by summing participants' responses across the nine items ($\alpha = .79$).

Demographic information. Participants' sex and previous formal performance rating experience were used as control variables. These variables were captured in a single question each: "what is your gender?" and "do you have any previous experience formally evaluating others?", respectively.

Cronbach's four components of accuracy. Rater accuracy can be defined in various ways. Cronbach (1955) identified four components of accuracy:

1) Elevation is the differential grand mean. In the current study, *elevation* represents each rater's mean ratings across performance dimensions (e.g., oral communication) and rating targets as compared to the mean ratings made by all other raters in that rating condition (i.e., DA/TA). As interrater reliability increases, the elevation of scores decreases.

2) Differential Elevation is used to assess the differential main effect of stimuli – that is, *differential elevation* is used to assess raters' agreement regarding each rating target. It is represented here by each participant's ratings (across

dimensions) of each rating target in comparison to the mean rating of all other participants in that rating condition (i.e., DA/TA).

3) Stereotype Accuracy is similar to *differential elevation* but the focus of *stereotype accuracy* is on assessing raters' agreement regarding each dimension, rather than each rating target. Here, *stereotype accuracy* represents the difference between each rater's appraisals on each dimension (across ratees) as compared to the mean ratings put forth by all participants in the same condition.

4) Differential Accuracy represents the interaction of the differential effects of the ratee and the dimension. Rather than focusing on main effects, *differential accuracy* evaluates the interaction between a) different rating targets and b) different skill dimension. Here, raters' scores on these combinations are compared to the mean ratings for that rating condition (i.e., DA/TA) for each of the combinations.

Lower scores on each of these measures reflect less variability in ratings, and thus, higher agreement among participants. In addition, due to the different aspects of interrater agreement each component reflects (e.g, agreement across targets; agreement across skill dimensions), each component score for each participant is necessarily aggregated across targets.

True scores. An additional way to assess accuracy is to compare ratings to a standard, or "true" score. True scores for each target's performance were also developed separately for DA and TA. This was due to the differing natures of these two different types of rater response formats.

Separate expert panels were used in creating DA and TA true scores. Expert panels were made up of certified assessment center raters currently serving as assessors in the university psychology department's assessment center program. Raters developed ratings independently before meeting to discuss their ratings, and then came to consensus upon a single true score for each target performer in regard to each rating dimension (e.g., oral communication, teamwork).

To ensure that DA and TA true scores were comparable, corresponding true scores were required to fall within .49 points of one another. For example, if a TA true score was "2", the corresponding DA true score must fall between 1.51 and 2.49. In the instances where DA and TA true scores differed by more than .49 points raters reconvened to discuss ratings until true scores fell within the acceptable range.

RESULTS

Descriptive Statistics

Means and standard deviations for the variables examined are depicted separately for TA and DA users in Table 1. Overall, it appears that DA users agreed more than did TA users. The only exception was in the case of differential elevation agreement for inconsistently performing rating targets. At this point it should be mentioned that lower means associated with Cronbach's four components of accuracy suggest less disagreement (i.e., higher rater agreement). In regards to true score accuracy, both TA and DA users appear to have rated target performance fairly accurately with both groups having mean true score differences of less than half of a point. However, it does appear that there was less variability in DA users' ratings. Finally, Table 1 also suggests that DA users experienced significantly higher cognitive load as means for both cognitive effort and cognitive difficulty were higher in the DA condition.

Table 2 provides the correlations of these variables. The correlations between response format (i.e., DA/TA) and each of Cronbach's four components of accuracy range from low to moderate (-.09 to -.44) in the direction suggesting higher agreement among DA users. The positive direction of the correlations between response format and the two cognitive load measures (.12 and .17) suggest that DA users experienced more cognitive load than TA users. Finally, the correlation between response format and true score difference is almost zero.

There were a number of other correlations that also may be interest. First, there were low to moderate (.07 to .35) correlations between Cronbach's four components of accuracy, and a moderate correlation between the two cognitive load measures(.39). Finally, true score difference moderately correlated with each of Cronbach's four components (.23 to .45), but the same was not true with regards to the cognitive load measures (.06 and .08).

Table 1. Descriptives for TA and DA rating conditions.

	TA (N=97)		DA (N=104)	
	M	SD	M	SD
Cognitive Effort	4.73	1.85	5.37	1.79
Cognitive Difficulty	3.16	1.37	3.51	1.50
Conscientiousness	4.05	0.49	4.00	0.52
Sex	0.72	0.45	0.71	0.46
Previous management experience	0.88	0.33	0.80	0.40
Elevation	0.19	0.14	0.17	0.13
Differential elevation	0.31	0.15	0.27	0.12
Stereotype accuracy	0.13	0.10	0.08	0.07
Differential accuracy	0.26	0.15	0.15	0.07
Elevation (i)	0.29	0.23	0.24	0.18
Differential elevation (i)	0.16	0.15	0.18	0.13
Stereotype accuracy (i)	0.19	0.19	0.12	0.10
Differential accuracy (i)	0.13	0.15	0.10	0.07
Elevation (c)	0.17	0.15	0.16	0.13
Differential elevation (c)	0.15	0.11	0.12	0.10
Stereotype accuracy (c)	0.18	0.12	0.10	0.08
Differential accuracy (c)	0.12	0.09	0.06	0.05
Average true score difference	0.38	0.19	0.38	0.13

(i) = agreement for inconsistently performing rating targets; (c) = agreement for consistently performing rating targets.

Table 2. Correlations.

	1	2	3	4	5	6	7	8	9	10	11
1. Mean Cog. Effort	-										
2. Mean Cog Difficulty	0.39	-									
3. Conscientiousness	0.17	-0.03	-								
4. Sex	-0.03	0.02	0.22	-							
5. Prev. Mgmt Exp.	-0.11	-0.08	-0.04	0.14	-						
6. Elevation	0.07	0.05	0.04	0.08	0.02	-					
7. Differential Elevation	0.06	-0.01	-0.10	-0.13	0.04	0.07	-				
8. Stereotype Accuracy	0.01	0.00	0.05	0.04	0.07	0.11	0.20	-			
9. Differential Accuracy	0.00	0.08	0.05	0.08	0.10	0.11	0.22	0.35	-		
10. Response Format	0.17	0.12	-0.05	-0.01	-0.11	-0.09	-0.15	-0.29	-0.44	-	
11. True Score Difference	0.08	0.06	-0.06	-0.13	0.15	0.23	0.45	0.23	0.41	-0.01	-

Hypotheses 1 and 2

Cronbach's components of accuracy. I examined the effect of response format (i.e., TA, DA) on Cronbach's four components of accuracy using a repeated multivariate or "doubly" MANOVA design in SPSS. This approach allows for the multiple dependent variables, each with repeated measurements (Leech, Caplovitz-Barrett, & Morgan, 2005). The multiple dependent variables in this analysis were each of Cronbach's four components of accuracy. Because each participant rated both inconsistent and consistent targets, target consistency was treated as a within-participant predictor. In addition, response format, sex, previous management experience, and conscientiousness were included as between-participant variables.

Only response format showed a significant multivariate effect, $F_{4, 193} = 9.73$, $p < .001$, $\eta^2 = .17$. No other between-participant predictors, target consistency (within-participant), nor any of the cross-level interactions were significant. Univariate findings suggest that response format had a significant effect on two of Cronbach's components: stereotype accuracy, $F_1 = 31.36$, $p < .001$, $\eta^2 = .14$, and differential accuracy, $F_1 = 22.30$, $p < .001$, $\eta^2 = .10$. Mean values for each of these two components were lower in the DA group than in the TA group (Table 2). As higher values indicate less agreement, this implies that DA users had higher agreement than TA users. This provides support for Hypothesis 1_a and is consistent with the findings of Fox et al (2002).

However, no support was found for Hypothesis 2_a which posited that the difference in rater agreement between TA and DA users will be more pronounced under the conditions of inconsistent target performance. This comes as a result of the non-significant interaction between target consistency and response format.

True score accuracy. I examined the effect of response format on true score accuracy using multilevel modeling (MLM) in statistical package R (R Core Development Team, 2011) to test Hypotheses 1_b and 2_b. MLM was used due to the nested structure of the true score accuracy data. That is, eight separate ratings were nested within each participant. This analysis included three dichotomous within-participant (Level 1) predictors: a) target consistency (inconsistent/consistent, b) target level (i.e., low/high), and c) skill dimension (communication/teamwork). These predictors reflect the differences in the performance of the four targets and the fact that participants made two ratings for each target. Between-participant (Level 2) predictors included response format, conscientiousness, sex, and previous management experience. Two dummy coded variables reflecting the order in which participants a) viewed rating targets, and b) rated skill dimensions were also included to test for ordering effects.

To determine which variables best predicted accuracy, I tested a series of models. The results of each of the five different models tested are depicted in Table 3. The log-likelihood values, reported as chi-square values, for each model were compared using the deviance test (ΔD) to determine whether or not predictors should be retained in subsequent models (Hox, 2002; Willet & Singer, 2003). Model 1 ($\chi^2_3 = 2206$) reflects the unconditional model – that is, a model free of any Level 1 or Level 2 predictors. The main purpose of such of a model is to identify the amount of variance existing within and between participants' ratings. The intraclass correlation (ICC) for this model suggests that less than 1% of the variance in rater accuracy was due to between-participant differences, and more than 99% was due to within-participant differences. Clearly, the vast majority of the variance found in this analysis lies within each participant's ratings.

Table 3. MLM predicting true score accuracy.

Model Fit	Model 1	Model 2		Model 3		Model 4		Model 5	
Residual	.00	.00		.00		.00		.00	
Intercept	.23	.18		.17		.18		.17	
Level 1 Predictors									
					**				
PL		-.29	***	-.25	*	-.29	***	-.43	***
PC		-.04		-.03		-.04		-.30	***
SD		-.15	***	-.11		-.15	***	-.31	***
Level 2 Predictors									
TORDER				-.01					
SORDER				.09					
CON						-.01			
SEX						-.02			
PMGMT						.00			
RF								-.28	***
Interactions									
PL x PC		.27	***	.24	*	.27	***	.51	***
PL x SD		.13	*	.04		.13	*	.22	*
PC x SD		-.11		-.10		-.11		.12	
PL x PC x SD		.63	***	.42	**	.63	***	.63	***
PL x TORDER				-.01					
PC x TORDER				.01					
SD x TORDER				-.03					
PL x SORDER				-.07					
PC x SORDER				-.05					
SD x SORDER				-.06					
PL x PC x SD									
PL x PC x TORDER				.06					
PL x SD x TORDER				.11					
PC x SD x TORDER				-.03					
PL x PC x SORDER				.01					
PL x SD x SORDER				.08					
PC x SD x SORDER				.02					
PL x PC x SD x TORDER				.18					
PL x PC x SD x SORDER				.30					
SD x RF								.31	***

PL x RF	.29	***
PC x RF	0.50	***
PL x SD x RF	-0.17	
PC x SD x RF	-0.45	***
PL x PC x RF	-0.45	***
PL x PC x SD x RF	0.01	

PL = Performance Level (i.e., low/high), PC = Performance Consistency (i.e., consistent/inconsistent), SD = Skill Dimension (i.e., communication/teamwork), TORDER = the order in which participants rated targets, SORDER = the order in which participants rated skill dimensions, CON = Conscientiousness, SEX = Participant sex, PMGMT = Previous management experience, RF = Response format.

In other words, rater accuracy was influenced more by characteristics of the specific rating target than by characteristics of the rater, including response format.

Level 1 (i.e., within-participant) variables, which based on the ICC analysis accounted for 99% of the variance in the dataset, were added in Model 2 (χ^2_{10} = 1805.60). Model 2 showed a significantly better fit than Model 1, ΔD (7) = 400.40, $p < .001$. As shown in Table 3 (under Model 2) a significant main effect was found for both skill dimension, β = -.15, t = -3.58, $p < .001$, and target level, β = -.29, t = -6.82, p < .001. A non-significant main effect was found for target consistency. However, each of the Level 1 predictors were part of significant interaction effects (see Table 3 under Model 2), including the three-way interaction between all Level 1 predictors, β = .63, t = 7.56, $p <$ .001. Thus, all three Level 1 main effects and significant interactions in Model 2 were retained for subsequent models.

To test for potential confounding effects due to the order in which participants viewed the targets or rated the dimensions, Level 3 added dummy-coded variables reflecting these ordering differences to the Level 1 predictors and interactions (χ^2_{26} = 1800). The deviance test indicated that Model 3 provided a non-significant improvement in fit over Model 2, ΔD (16) = 5.60, n.s. In addition, neither of the ordering main effects

nor any of their interactions were significant (see Table 3 under Model 3). Thus for the sake of parsimony, neither of the ordering main effects nor their interactions were retained for subsequent analyses.

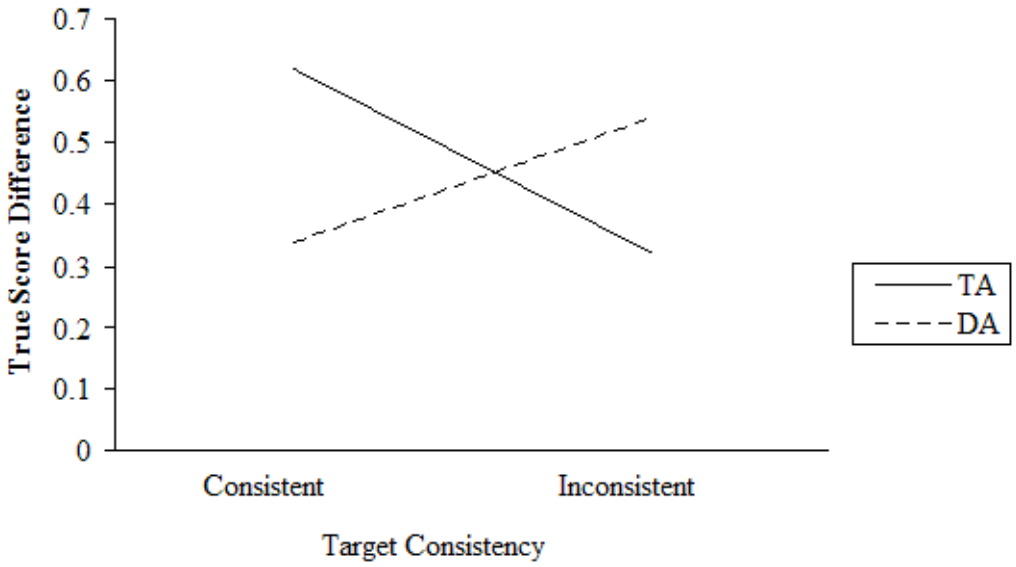
Model 4 included the Level 1 main effects and interactions from Model 2 and added the main effects for the Level 2 covariates (i.e., conscientiousness, sex, and previous management experience) ($\chi^2_{13} = 1821$). Model 4 fit the data worse than Model 2, ΔD (3) = -15.4, n.s., and none of the newly added main effects were significant. Thus, none of the Level 2 covariates examined in Model 4 were included in the final model.

Model 5 ($\chi^2_{18} = 1773$) introduced the main effect of the response format and its cross-level interactions with Level 1 predictors. Model 5 fit the data significantly better than Model 2, ΔD (8) = 32.60, $p < .001$. There were significant main effects for all three Level 1 predictors (see Table 3 under Model 5) and for response format, $\beta = -.28$, $t = -4.83$, $p < .001$. Specifically, the DA format was associated with greater accuracy, providing support for Hypothesis 2_a.

There were a number of significant interactions among the predictors in Model 5. The most relevant of these was the significant 2-way interaction found between target consistency and response format, $\beta = .50$, $t = 6.05$, $p < .001$. This interaction is important for two reasons. First, it sheds light on the main effect found for response format. Although a main effect was found for response format, Figure 5 suggests that DA users were only more accurate than TA users when target performance was consistent. Related to this, this interaction directly relates to Hypothesis 2_b and suggests that TA, and not DA, ratings were more accurate in regards to inconsistently performing targets. Thus,

Hypothesis 2_b was not supported, as the interaction was significant, but in the opposite direction.

Figure 5 – Interaction between Performance Consistency and Response Format Type in Predicting True Score Accuracy.



Target performance level and response format also significantly interacted (β = .29, t = 3.46, p = .001) as DA was more accurate than TA when target performance level was low (Figure 6). A similar result was found for skill dimension and response format (β = .31, t = 3.74, p < .001) suggesting that DA users were more accurate in rating targets' communication skills than were TA users (Figure 7). Other significant 2-way interactions were found between Level 1 predictors. The interaction between target level and skill dimension (β = .22, t = 2.58, p = .01) suggests that rating accuracy was greater for rating communication skills when target level was low (Figure 8). The interaction between target level and target consistency (β = -.45, t = -3.87, p < .001) suggests that

rating accuracy was greater for low performers when target performance was consistent, but greater for high performers when target performance was inconsistent (Figure 9). All of the Level 1 interactions should be interpreted with caution, as they may reflect idiosyncratic characteristics of the four different targets.

Figure 6. Interaction between Performance Level and Response Format Type in Predicting True Score Accuracy.

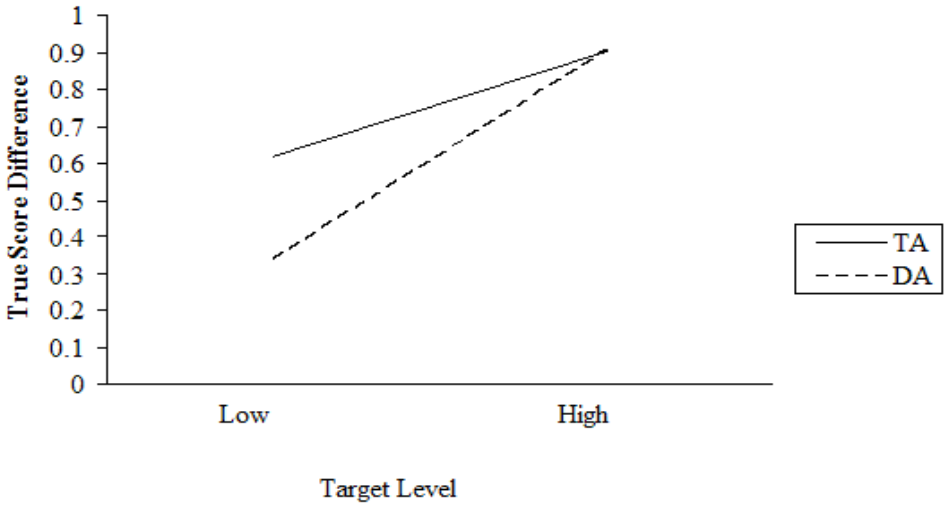


Figure 7. Interaction between Skill Dimension and Response Format Type in Predicting True Score Accuracy.

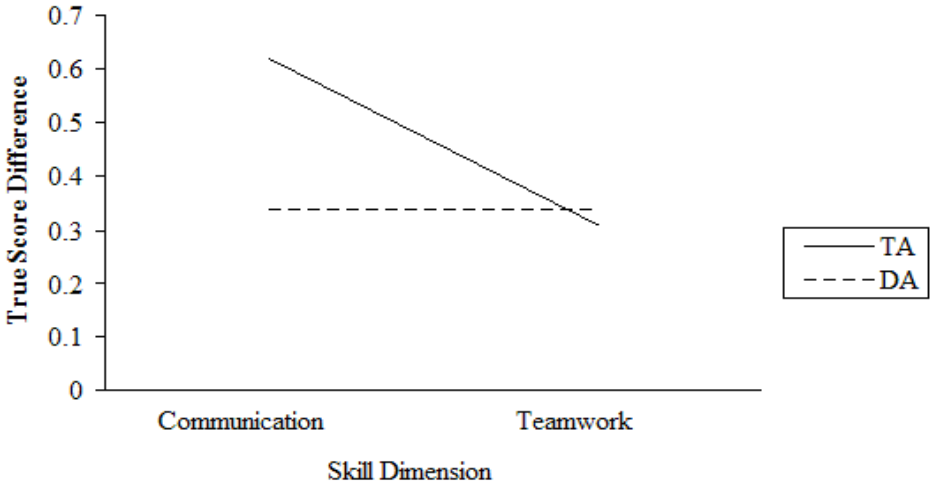


Figure 8. Interaction between Performance Level and Skill Dimension in Predicting True Score Accuracy.

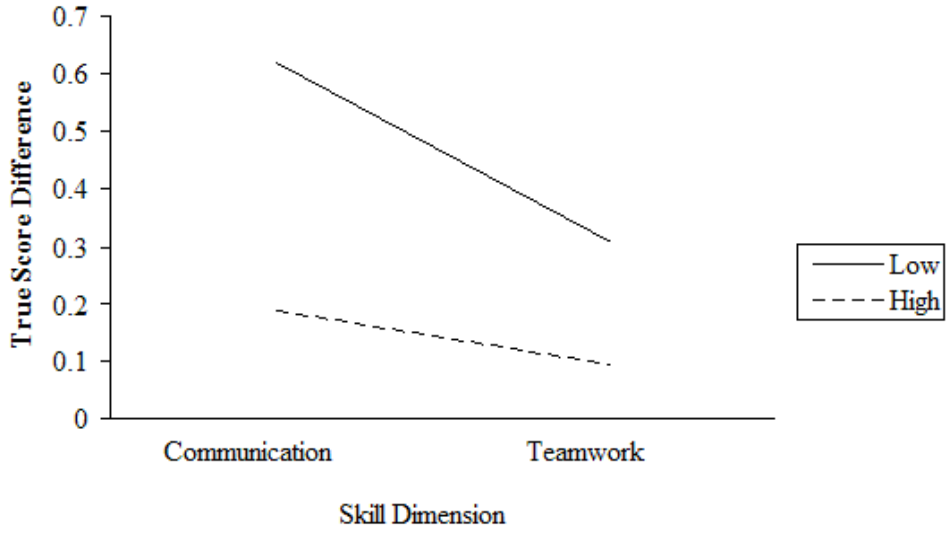
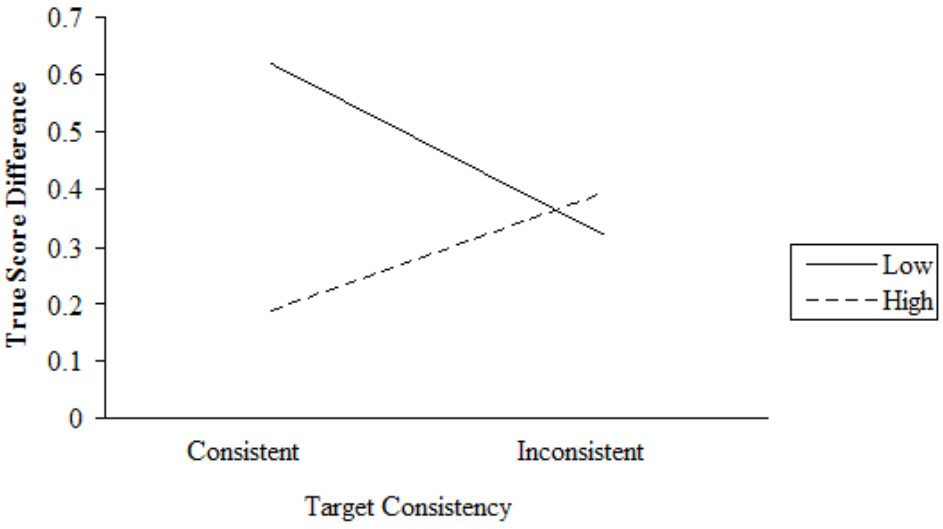


Figure 9. Interaction between Performance Level and Performance Consistency in Predicting True Score Accuracy.



In addition, to the 2-way interactions, there were also significant 3-way interactions. Target skill dimension, target consistency, and response format significantly interacted, $\beta = .51$, $t = 6.03$, $p < .001$. Figure 10 suggests that skill dimension did not interact with consistency to affect rating accuracy in the DA condition. On the other hand, skill dimension did interact with consistency in the TA condition, as rating accuracy of target communication remained the same between consistent and inconsistent targets, but was higher for inconsistent targets for teamwork.

Target consistency and response format also significantly interacted with target performance level, $\beta = -.45$, $t = -3.85$, $p < .001$. With regard high performers, raters in both DA and TA were less accurate in rating inconsistent performers than consistent performers. However, DA users' accuracy in rating low performers remained the same across target consistency conditions, but TA users were actually more accurate in rating the low/inconsistent target than the low/consistent target (Figure 11).

Figure 10. Three-way interaction between Performance Consistency, Skill Dimension, and Response Format Type in Predicting True Score Accuracy.

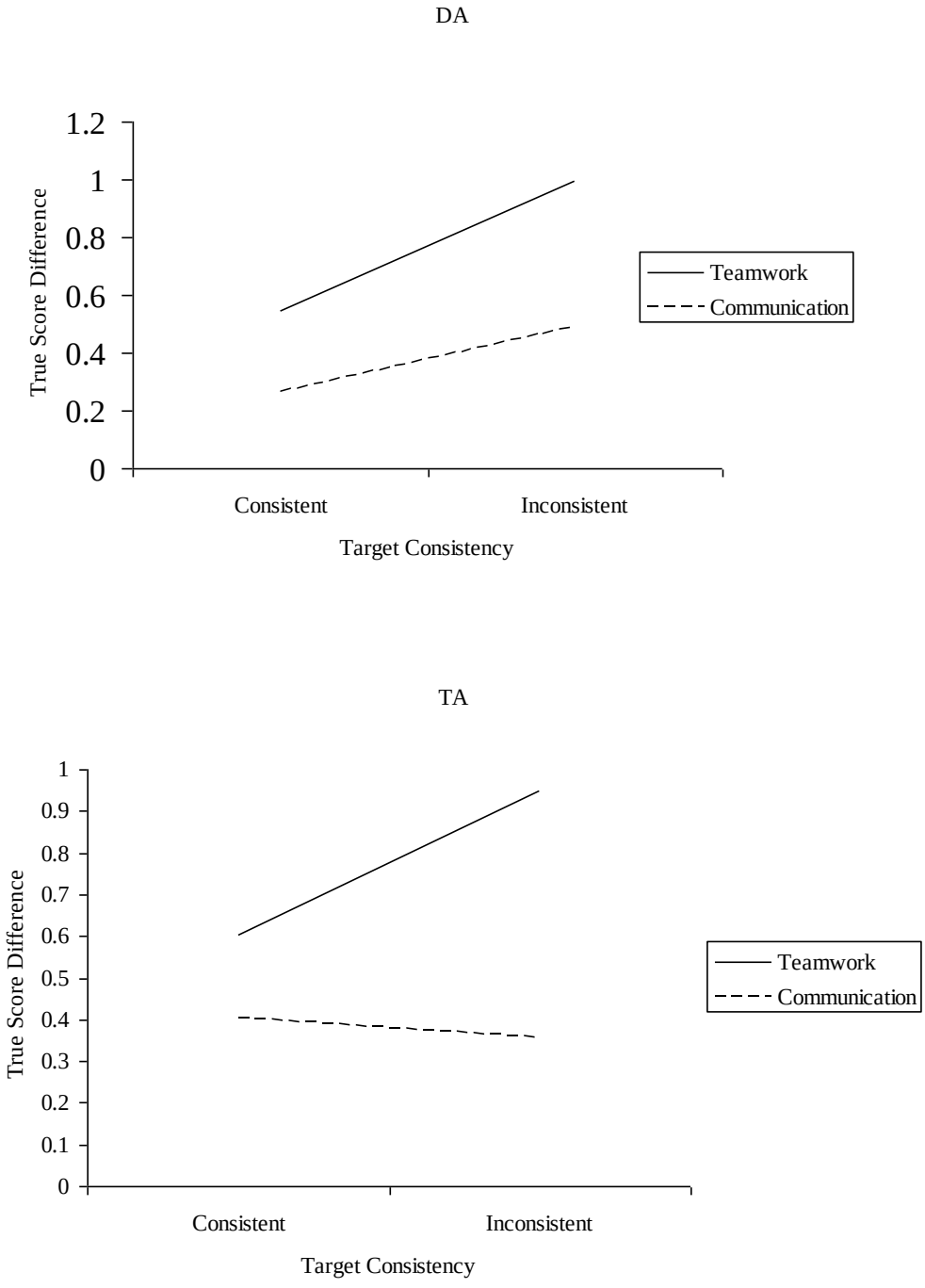
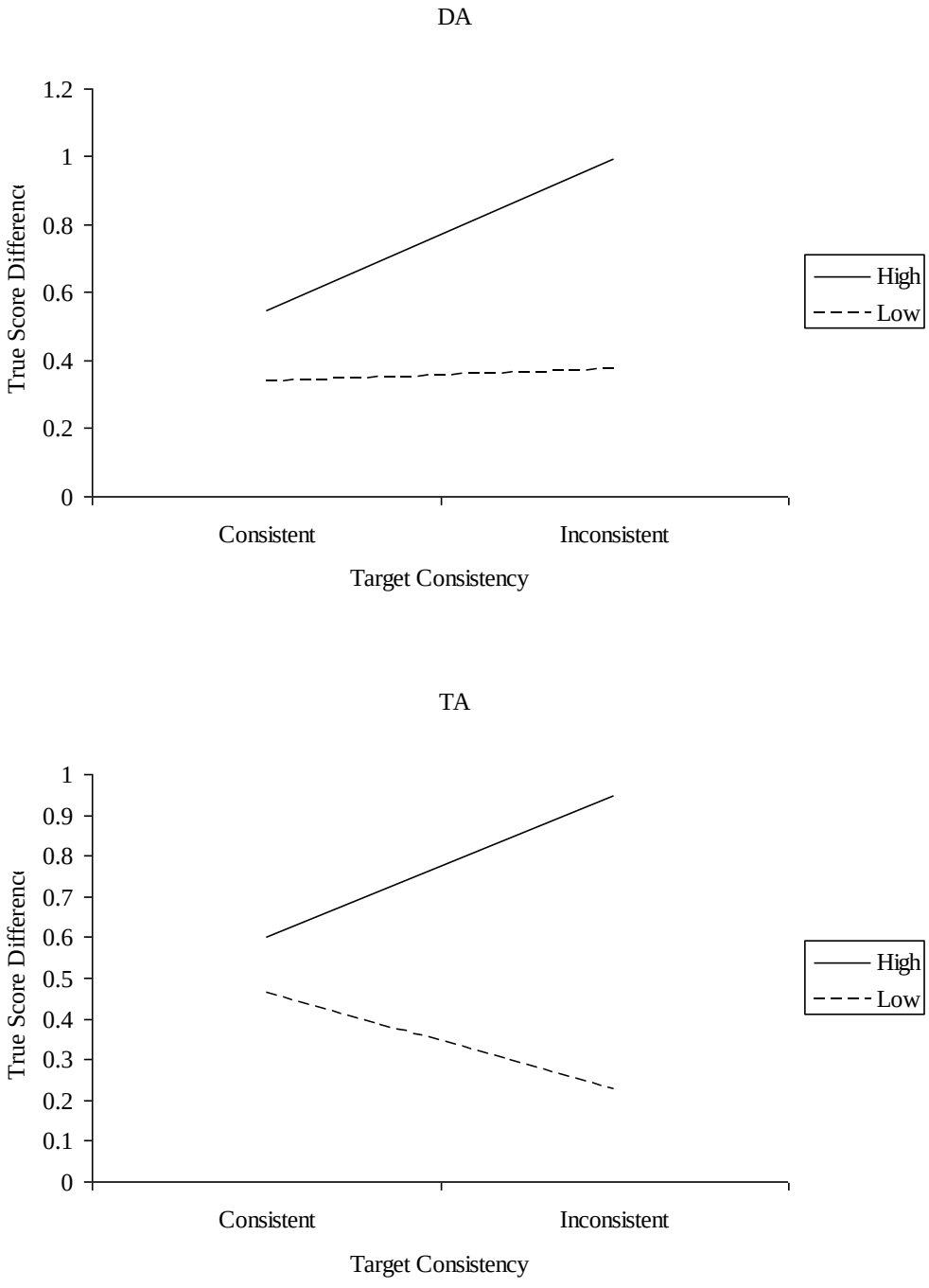


Figure 11. Three-way interaction between Performance Consistency, Performance Level, and Response Format Type in Predicting True Score Accuracy.



Hypothesis 3

Hypothesis 3 used the same multilevel framework as the analyses for true score accuracy. To test Hypothesis 3, the effect of response format was examined in relation to two separate self-reported measures of cognitive load – cognitive effort and cognitive difficulty. Findings are presented below, but it is important to first point out a major difference between the analyses for Hypothesis 3 and the true score accuracy analysis used to examine Hypotheses 1_b and 2_b. That is, participants only made one cognitive effort and cognitive difficulty rating for each target. Thus, only the Level 1 predictors of performance consistency and level, and not skill dimension, were included in the analyses to test Hypothesis 3.

Cognitive effort. Findings from each of the models described below are available in Table 4. Model 1 ($\chi^2_3 = 5146$) resulted in an ICC of .78. This suggests that 78% of the variance in participants' cognitive effort lies in differences between participants, and the remaining 22% lies in differences within participants.

The main effects of both Level 1 predictors (target consistency and level), and their interaction were added in Model 2 ($\chi^2_6 = 5118$). Model 2 showed a significantly better fit than Model 1, ΔD (3) = 28, $p < .001$. However, only the main effect of target consistency was significant, $\beta = .16$, $t = 2.43$, $p < .05$, suggesting that inconsistent targets required more effort to rate. Thus, only the main effect of target consistency was retained in subsequent models.

Table 4. Multilevel Models for Predicting Cognitive Effort.

Model Fit	Model 1	Model 2		Model 3		Model 4	
Residual	.94	.92		.93		.92	
Intercept	3.28	3.29		3.09		3.11	
Level 1 Predictors							
PC		.16	*	.25	***	.41	***
PL		.07					
Level 2 Predictors							
TORDER				-.39			
SORDER				-.07			
CON				.73	**	.32	
PMGMT				-.38			
SEX				-.25			
RF				.68	**	.83	**
Interactions							
CON x RF						.56	
PC x RF						-.32	***
PC x CON						.13	
PC x CON x RF						-.11	

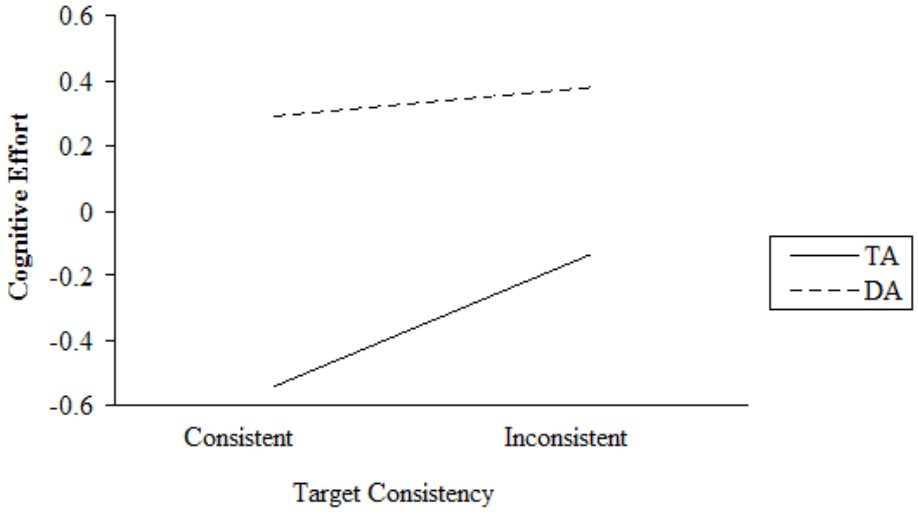
PL = Performance Level (i.e., low/high), PC = Performance consistency (i.e., consistent/inconsistent), TORDER = order in which participants rated targets, SORDER = order in which participants rated skill dimensions, CON = conscienciousness, PMGMT = whether or not participants had any previous formal performance rating experience, SEX = participant sex, RF = response format type (i.e., TA/DA).

Model 3 ($\chi^2_{10} = 5110$) introduced the main effects of all Level 2 variables including both ordering effect variables, conscientiousness, sex, previous management experience, and response format. Model 3 showed only a marginally better fit over Model 2, ΔD (4) = 8, $p = .09$. However, the main effects of conscientiousness ($\beta = .73$, $t = 2.83$, $p < .01$) and response format ($\beta = .68$, $t = 2.65$, $p < .01$) were significant. Moreover, performance consistency remained significant, $\beta = .25$, $t = 5.131$, $p < .001$. Despite the fact that the fit of Model 3 was only marginally better than the previous

model, response format and conscientiousness were retained along with target consistency in the final model in order to examine potential interaction effects.

Model 4 ($\chi^2_{10} = 5106$) added the interaction effects between the predictors retained from Model 3. Model 4 showed significantly better fit than Model 2, ΔD (4) = 12, $p <$ .05. There was a significant main effect for target consistency, $\beta = .41$, $t = 5.93$, $p < .001$. This suggests that rating inconsistent targets required more cognitive effort from participants in both the TA and DA conditions. A significant main effect was also found for response format, $\beta = .83$, $t = 3.21$, $p < .01$. However, this finding was in the opposite direction than was hypothesized and suggests that DA users actually reported the rating tasks to require more cognitive effort than TA users. Finally, the significant cross-level interaction between target consistency and response format type ($\beta = -.32$, $t = -3.31$, $p =$ .001) is shown in Figure 12 and suggests that, although DA users reported greater cognitive effort in rating both levels of target consistency, that difference was greater when target performance was consistent.

Figure 12. Interaction between Performance Consistency and Response Format Type in Predicting Cognitive Effort.



Cognitive difficulty. Each of the models described below are shown in Table 5. Model 1 ($\chi^2_3 = 5240$) – that is, the unconditional model – resulted in an ICC of .64, suggesting that approximately 64% of the variance in the data was associated with differences between participants' reported levels of cognitive difficulty, and the remaining 36% was associated with differences within participants.

Table 5. Multilevel Models Predicting Cognitive Difficulty.

Model Fit	Model 1	Model 2		Model 3		Model 4	
Residual	1.95	1.96		1.98		1.94	
Intercept	1.08	0.99		0.99		0.99	
Level 1 Predictors							
PL		-0.14	*	-0.14	*	0.09	
PC		0.22	**	0.22	**	0.40	***
Level 2 Predictors							
TORDER				-0.11			
SORDER				0.20			
CON				-0.12			
PMGMT				-0.36			
SEX				0.23			
RF						0.67	**
Interactions							
PL x RF						-0.45	**
PC x RF						-0.35	*
PL x PC		0.52	***	0.52	***	0.34	*
PL x PC x RF						0.34	

PL = Performance Level (i.e., low/high), PC = Performance consistency (i.e., consistent/inconsistent), TORDER = order in which participants rated targets, SORDER = order in which participants rated skill dimensions, CON = conscienciousness, PMGMT = whether or not participants had any previous formal performance rating experience, SEX = participant sex, RF = response format type (i.e., TA/DA).

Model 2 ($\chi^2_6 = 5130$), which again included the two Level 1 predictors and their interaction, showed significantly better fit than Model 1, ΔD (3) = 110, $p < .001$. The main effect of both target consistency and level significantly predicted participant-reported cognitive difficulty, $\beta = .22$, $t = 3.11$, $p < .01$, and $\beta = -.14$, $t = -1.98$, $p < .05$, respectively. The interaction was also significant, $\beta = .52$, $t = 5.20$, $p < .001$. Thus, all three predictors were retained in subsequent models.

Model 3 introduced the main effects of the Level 2 covariates ($\chi^2_{11} = 5132$), and showed worse fit than Model 2. Both Level 1 predictors remained significant (see Table 5 under Model 3). None of the Level 2 main effects were significant. Thus, the Level 2 covariates in Model 3 were not retained in the final model.

Model 4 ($\chi^2_{10} = 5122$) included the main effects of target consistency and level, and introduced the main effect of response format, as well as all possible interactions between these three predictors. However, the fit of Model 4 was only marginally better than Model 2, ΔD (4) = 8, $p < .09$. Thus, it appears that adding the main effect of response format and its interactions with the Level 1 predictors resulted in a non-significantly better model fit.

Nonetheless, there were significant main effects for target consistency ($\beta = .40$, $t = 3.99$, $p < .001$) and response format ($B = .67$, $t = 3.02$, $p < .01$). These findings are similar to those for cognitive effort in that all participants found rating inconsistent performers to be more difficult, and it was again the DA users, and not TA users, who reported the rating tasks to be more difficult. In addition, there were significant interactions between response format type and target consistency ($\beta = -.35$, $t = -2.53$, $p < .05$) and level ($\beta = -.45$, $t = -3.20$, $p < .01$). These interactions suggest that, although DA users reported the rating task to be more difficult across all rating targets, greater differences existed between these groups when target performance was low and when target performance consistent, respectively. These interactions are depicted in Figures 13 and 14. Finally, target consistency significantly interacted with target level ($\beta = .34$, $t = 2.39$, $p < .05$), suggesting there were greater differences in cognitive difficulty between

consistent and inconsistent performers when target performance was high than when target performance was low (Figure 15).

Figure 13. Interaction between Performance Consistency and Response Format Type in Predicting Cognitive Difficulty.

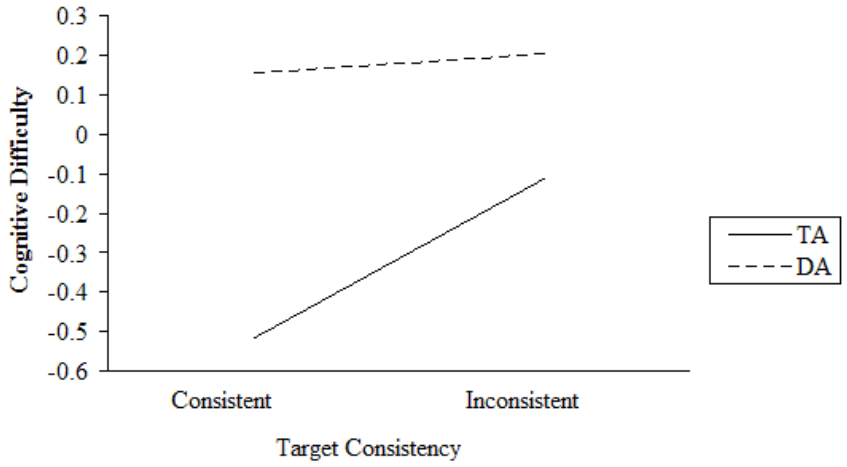


Figure 14. Interaction between Performance Level and Response Format Type in Predicting Cognitive Difficulty.

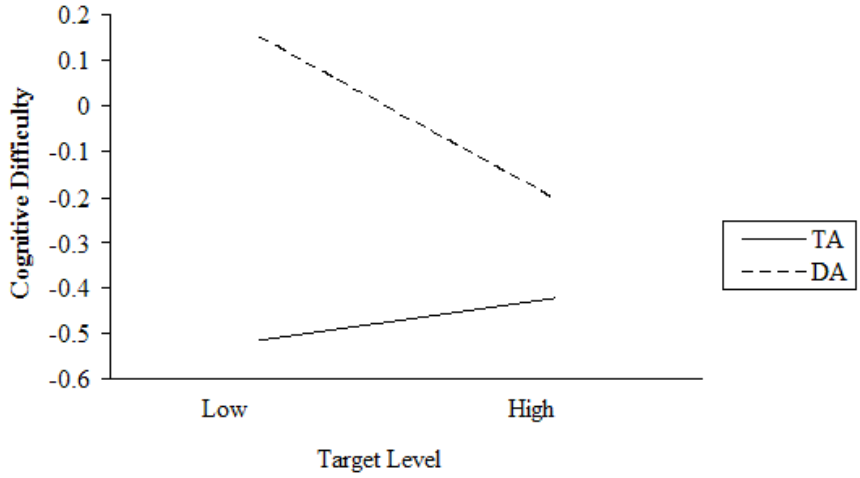
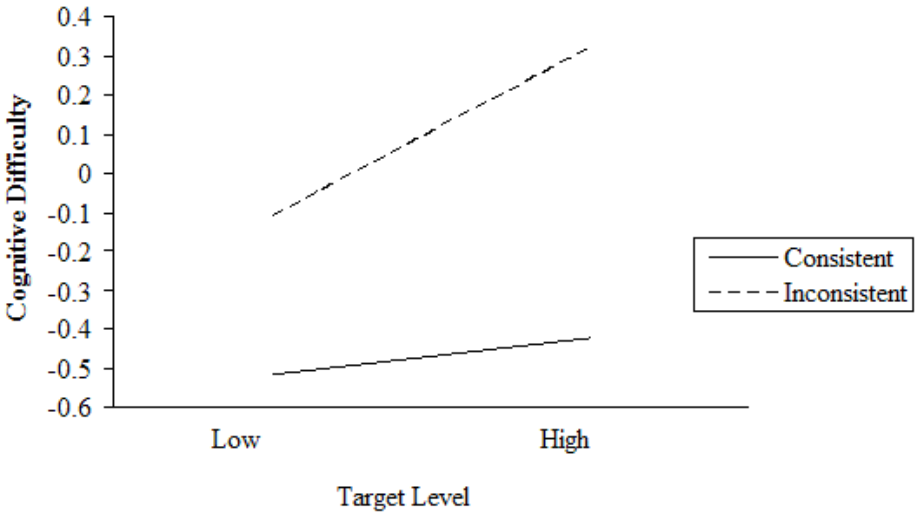


Figure 15. Interaction between Performance Level and Performance Consistency in Predicting Cognitive Difficulty.



Previous research suggests that one potential reason for increased accuracy on the part of DA users is the result of a reduction in experienced cognitive load (e.g., Kane, 2000). Findings regarding the two cognitive load measures strongly suggest that this is not the case, but instead that DA users experience more cognitive load during the rating task. Indeed, it appears that the increased accuracy experienced by DA users may be the result of more cognitive resources being put towards the rating task. Thus, Hypothesis 3 was not supported.

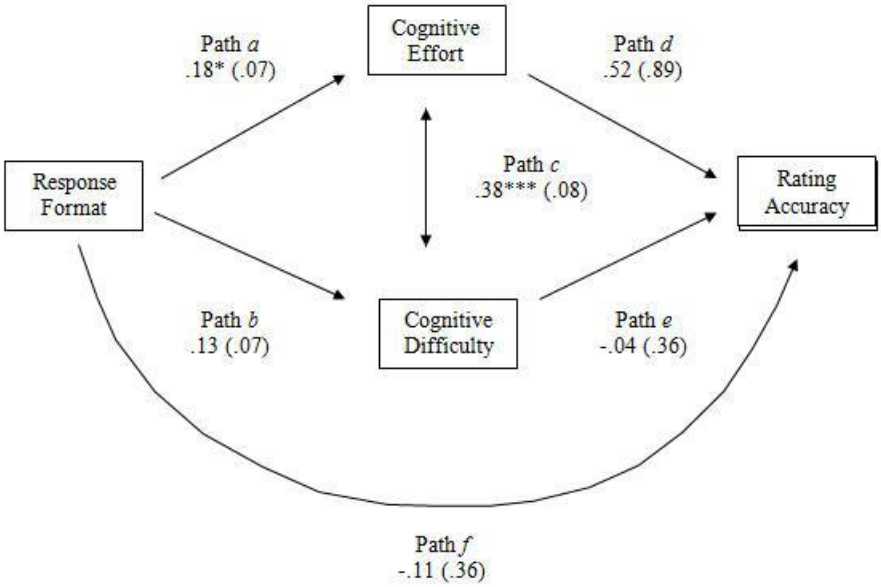
Hypothesis 4

The final hypothesis proposed that cognitive load mediated the relationship between the response format used and rater accuracy. To test this, I used the multilevel mediational approach using SEM (LeBreton, Wu, & Bing, 2008; Preacher, Zyphur, & Zhang, 2010) . The SEM mediational model was examined in Mplus (Muthen & Muthen, 2007). The mediational model tested here was theoretically a three level model (with

rater accuracy for each skill dimension nested within overall rating accuracy which, in turn, was nested within response format). Mplus is not currently equipped to handle three level models. Thus, ratings for communication and teamwork were analyzed separately. In order to avoid increasing the family-wise error rate, the significance level was reduced to .025 for the mediational paths from cognitive effort and cognitive difficulty to rater accuracy and the direct path from response format type to rater accuracy. The .05 significance level was used for the paths from response format type to cognitive effort and cognitive difficulty because they were not affected by which dimension was rated, and thus, were identical across the two analyses (see Figures 16 and 17). Intraclass correlation (ICC) values in both of the following analyses reflect the amount of variance in each outcome variable in the models that is accounted for by the type of response format used by participants.

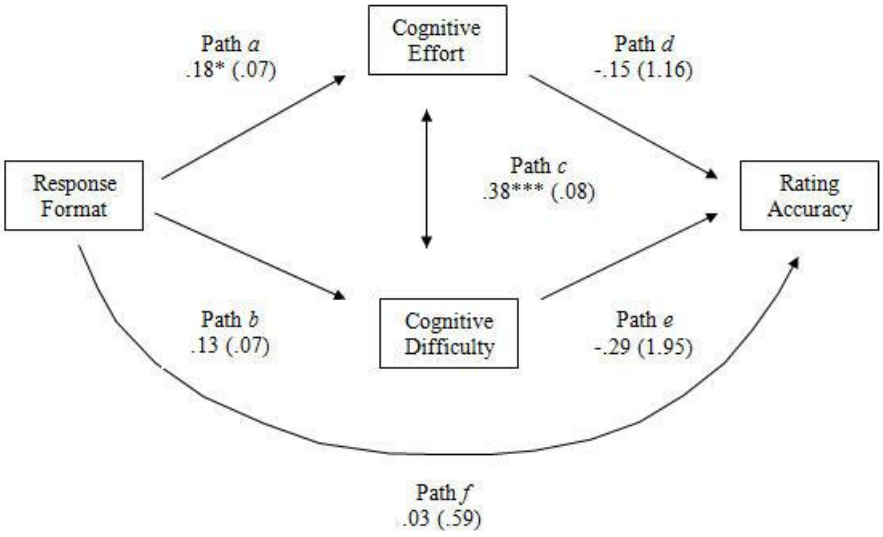
In the model predicting accuracy in communication ratings ($\chi^2_9 = 6414.50$), the ICC values indicate that response format accounted for 74% of the variance in cognitive effort, 58% of the variance in cognitive difficulty, and 2% of the variance in rating accuracy. This suggests that the remaining 26%, 42%, and 98% of the variance in cognitive effort, cognitive difficulty, and rater accuracy, respectively, varies within participants. This model showed a significant, positive relationship between response format and cognitive effort, $\beta = .18$, $p < .05$ (path *a*), suggesting that participants who used the distributional response format reported using more effort during the rating task. There was also a significant, positive relationship between cognitive effort and cognitive difficulty, $\beta = -.38$, $p < .001$ (path *c*). None of the remaining paths were found to be significant (Figure 16).

Figure 16. Multilevel Mediation SEM Model for Communication Skill Dimension.



Standard errors are in parentheses.

Figure 17. Multilevel Mediation SEM Model for Teamwork Skill Dimension.



Standard errors are in parentheses.

In the model predicting accuracy in teamwork ratings ($\chi^2_9 = 6855.02$), the ICC values indicate that 74% of variance in cognitive effort, 59% of the variance in cognitive difficulty, and 0.1% of the variance in rating accuracy was associated with response format. Thus, 26%, 41%, and 99.9% of the variance in cognitive effort, cognitive difficulty, and rating accuracy, respectively, was associated with differences within participants. As mentioned above, the paths from response format to cognitive effort and cognitive difficulty are necessarily identical to those in the communication model. And as in the communication model, neither paths leading from response format nor either of the cognitive load measures to teamwork rating accuracy were significant (Figure 17).

Taken together, the mediational analyses provide no support for Hypothesis 4.

DISCUSSION

The majority of previously published research has concluded that DA, in comparison to TA, increases rater accuracy (e.g., Edwards & Woehr, 2007; Fox et al., 2005; Kane, 2000). However, past research examining performance rating accuracy has focused almost exclusively on interrater agreement and reliability measures as accuracy criteria (e.g., Fox et al., 2005; Jako & Murphy, 1990; Steiner et al., 1993; Woehr & Miller, 1997). Findings from this study support the previous research suggesting that interrater agreement is higher among DA users than TA users. Specifically, in terms of Cronbach's components of stereotype accuracy and differential accuracy (Fox et al., 2005). In addition, this study also provides the first evidence that DA may lead to greater true score accuracy. However, this advantage appears to be slight. In addition, the finding that TA users were more accurate raters of inconsistent performance is in contrast to past findings (i.e., Deadrick & Gardener, 1997; Steiner et al., 1993).

Proponents have argued that one reason DA may lead to greater rater accuracy, or at least greater interrater agreement, is the result of a reduced level of cognitive load experienced by raters during the rating process (e.g., Edwards & Woehr, 2007; Kane, 2000). In this study, I directly tested this assumption by examining the effect of response format on two separate, self-reported measures of cognitive load, cognitive effort and cognitive difficulty. Interestingly, my findings provided no support for this assumption. In fact, they suggest that DA users put forth significantly more cognitive effort than their

TA counterparts, and DA users found the rating task to be more cognitively difficult as well.

I also tested the mediational hypothesis positing that the response format used by participants would lead to different levels of cognitive load, which, in turn, would lead to differences in rating accuracy. However, both SEM analyses resulted in largely null findings. Nonetheless, the results from this study shed some interesting light on the value of DA as a viable alternative to the way organizations commonly evaluate individual performance.

Interrater Agreement

In Hypothesis 1a, I proposed that DA would, in general, elicit greater interrater agreement than TA. As mentioned, this has been the focus of much of the previous research on DA, and the doubly MANOVA findings from this study are largely in line with past research. In fact, they mirror those of Fox et al. (2005), who also used Cronbach's four components of accuracy, by finding significantly greater agreement between DA users in terms of stereotype accuracy and differential accuracy. However, this study took the advantages of DA in terms of interrater agreement a step further by hypothesizing that it would be especially advantageous under conditions of inconsistent target performance. This hypothesis was based on findings by Deadrick and Gardner (1997) and the propositions made by both Kane (2000) and Kane and Woehr (2007) who suggested just that. Interestingly, the doubly MANOVA findings from this study did not suggest this to be the case, as I found no significant interaction between target consistency and response format for interrater agreement. Thus, findings from this study support much of the previous research regarding DA's advantage in interrater agreement,

in general, but do not support the idea that DA is especially advantageous under conditions of inconsistent performance.

True Score Accuracy

Interrater agreement measures such as Cronbach's components of accuracy can provide initial evidence for rater accuracy, but they are not true measures of it. A shortcoming of much of the previous literature is the lack of evidence suggesting differences between DA and TA in terms of actual rater accuracy. Thus, a major contribution of this study was to examine this by using expert panel-developed true scores.

Hypothesis 1b posited that DA users would have higher true score accuracy than TA users, and I found support for this hypothesis through the statistically significant difference favoring DA rater accuracy over TA rater accuracy. That said, it is important to treat this finding with caution considering that less than 1% of the variance in rater accuracy was due to between-participant differences. In other words, at most, there lies only less than 1% of the variability in ratings that was due to the response format used, suggesting that the practical difference in rating accuracy between DA and TA users is minimal. In addition, this main effect for response format (used to assess Hypothesis 1b) must be interpreted through the interaction found between response format and performance consistency, which suggests that TA users, and not DA users were actually more accurate when target performance was inconsistent (see Figure 5).

The minimal practical significance associated with my findings may have been a result of characteristics of the rating task in this study. The videos of target performance were each approximately eight minutes in duration. This in contrast to performance

ratings reflecting much longer appraisal periods in past research (e.g., 1 week durations in Deadrick & Gardener, 1997, and no less than four months in Fox et al., 2005).

In addition this study included rater training for all participants, a characteristic shown to have a strong effect on rater accuracy (e.g., Woehr & Huffcutt, 1994) which was also reflected in this data as all raters were fairly accurate. In addition, rater training not included in past DA research. Thus, it is entirely possible that TA raters were simply able to accurately rate target performance under these conditions. That said, performance evaluations that take place in the real world often reflect much longer periods of time (e.g., one year) and include much more precise response scales than the 4-point scale used here (an issue discussed below). Thus, although TA users may have been able to accurately rate target performance under the conditions of this study, this may not be the case in more complex rating situations.

Regardless of the potential reasons behind why there was little practical difference between TA and DA raters, this study contributed to the literature by examining both TA and DA in regards to true score accuracy. These findings do suggest a statistical difference favoring DA (Hypothesis 1_b). However, findings suggest that it is TA, and not DA, that better captures true performance under conditions of inconsistent performance.

Hypothesis 2b posited that the rating accuracy advantage of DA would be especially prominent under conditions of inconsistent performance. However, as noted above, this was not the case, and the interaction found between response format and target consistency suggested that, although DA users were more accurate when target

performance was consistent, the opposite was true when target performance was inconsistent.

That said, this finding may have had more to do with a lack of measurement precision due to the response scale used than it did with real differences between DA and TA in terms of rater accuracy. The rating scale I used for both the TA and DA condition was a 4-point scale. Previous research that has found differences between DA and TA have traditionally used more precise scales. For example, both Fox et al. (2005) and Woehr and Miller (1997) used 7-point scales. Deadrick and Gardner (1997), who found DA to be especially advantageous under conditions of inconsistent performance, used an 8-point scale.

Given that targets' performance was either high or low, the potential responses were further restricted to the point that TA users likely had a 50 percent chance of choosing the true level of performance – that is, either "1" or "2" in the case of low target performance or "3" or "4" in the case of high target performance. In contrast, DA users' mean ratings were not restricted to whole numbers as they were calculated from the distributional information they provided. Thus, there was a greater range of possible performance ratings (e.g., anywhere between "1" and "2.4" for low target performance). In general, this increased level of precision associated with DA should be seen as an advantage, regardless of the range of the rating scale. However, when a fairly restricted scale, such as a 4-point scale, is used to compare rater accuracy, the inherent differences between TA and DA favor TA and may reduce differences in accuracy between the two.

This restriction of range advantage may have been especially advantageous for TA users under inconsistent target performance conditions. For example, even when

target performance was inconsistent the true score associated with that performance was still within the "1" to "2" range making an accurate rating of that performance using the TA scale fairly easy. However, this may not have been the case for DA raters who were required to accurately record the variation in performance in order to make an accurate overall rating. Given this and that findings from this study suggest that accurately rating inconsistent performers is more difficult than accurately rating consistent performers (see Figure 13), the potential that TA users were advantaged simply due to characteristics of the scale may have led the significant interaction that was found. One finding from this study with potentially important implications is that, despite DA having the potential advantage of being more precise and thus increasing the chance of increasing variability among raters, is that DA users did show a greater level of overall interrater agreement than did TA raters. Thus, it would be particularly interesting to see whether or not this same interaction occurs when a more precise scale is used, such as the 7- or 8-point scale used in previous research (e.g., Deadrick & Gardener, 1997).

Cognitive Load and the Mediational Hypothesis

Hypothesis 3 was developed based on existing theory suggesting that DA users experience less cognitive load during the rating task and, as a result, are able to develop more accurate ratings (e.g., Edwards & Woehr, 2007; Kane, 2000). However, this assumption had not been empirically tested until now. What was actually found is that DA users experienced significantly *more* cognitive load during the rating task than did TA users. There are a number of possible explanations for these results. First, it may have been the result of the novelty of the DA response format. Few people are familiar

with this type of response format and, despite the training received prior to the rating tasks, this novel format may have required more cognitive resources from users.

Second, DA users may have experienced more cognitive load during the rating task because it requires that users specify frequencies at each performance level. This is in contrast to the TA format which simply requires users to mark a single mean rating. In theorizing why DA may result in a reduced cognitive load, Kane (2000) draws heavily on arguments by Hasher and Zacks (1984) regarding the automaticity of event frequency encoding. However, the way Hasher and Zacks use the term automatic does not suggest that the process is effortless. Instead, they use the term "automatic" in regards to humans having an innate tendency to think about differences in the frequency of events. Thus, supervisors' tendency to think about subordinates' performance in terms of frequencies likely still requires a significant amount of cognitive resources and simply having to provide more information through DA, in comparison to TA, may have resulted in an increased cognitive load.

A third possibility is that specifying this additional frequency information may have caused DA users to experience more cognitive load because they were forced to reflect more on targets' performance in justifying their ratings. A common argument against laboratory research is the low or no consequence associated with participation (e.g., Simonson & Nye, 1992). In this study, DA users, having reported higher levels of cognitive load, may have simply lost interest or the motivation to rate target accurately. In addition, there was no significant effect for conscientiousness in any of the analyses suggesting that DA users were no more motivated to perform than were TA users. This lends support to the idea that DA inherently forces users to reflect on target behavior, and

given the sample used here, this suggests that when DA is used in real world situations when users are motivated, it may prove to have a greater advantage over TA in terms of rater accuracy.

If DA users focus more of their information processing resources on accurately recalling and recording behavioral frequencies, we might expect that they should also be more accurate than their TA counterparts. This, of course, is opposite of what has been proposed in the previous literature. However, if DA requires more cognitive resources of users, this may imply that raters are more actively recalling behaviors rather than relying on heuristics, which may ultimately result in more accurate ratings.

This was tested in Hypothesis 4, and findings suggest that, regardless of the reason for the increased cognitive load, it did not mediate the relationship between response format type and true score accuracy. One potential reason for this finding is differences in participants' cognitive, or information processing ability. Findings from this study, although not supported through the mediational analysis, suggest that DA results in greater interrater agreement, true score accuracy, and cognitive load. In short, it may be plausible that individuals with higher cognitive processing ability are better able to handle the amount of cognitive load associated with using DA. In turn, these individuals may be better able to use the response format and provide more accurate ratings. On the other hand, individuals with low cognitive processing ability may find DA too difficult to use and their ratings may become less accurate.

Strengths of the Study

This study has a number of strengths, many of which were already discussed above in this section (e.g., providing rater training, measuring actual true score accuracy).

In addition, this study used a repeated-measures design that allowed a deeper investigation of the immediate relationship between cognitive load and ratings. For example, MLM was employed to examine both true score accuracy and the measures of cognitive load. In addition, multilevel SEM was used to examine the mediational hypothesis. Using the multilevel approach has a number of advantages when working with a nested dataset, including separating within- and between-participant variance and accurately estimating standard errors (e.g., Hox, 2002; Goldstein, 2011). Thus, this study used best practices in answering the questions raised through Hypotheses 1-4.

Limitations

As alluded to above, there was a potential lack of rating precision due to the use of a 4-point scale rather than, for example, a 7-point scale. This may have ultimately resulted in a restriction of range among ratings, and likely affected TA rating accuracy more than DA rating accuracy. This may have resulted in a restriction of range among true scores as well, and minimized real differences in rating accuracy between TA and DA raters. Future researchers on this topic should take the issue of measurement precision into consideration ensure comparable true scores across conditions and a better understanding of real differences between ratings using these two types of response formats.

Other potential limitations associated with this study are related to external validity. Although participants received training prior to the rating task, there are concerns regarding whether or not undergraduates have the same decision making ability and motivation as real world managers and assessment center assessors. That said, the motivation levels of undergraduate sample used here may not be that unrepresentative of

the real world supervisors. Indeed, researchers have largely assumed that the rating task is complex (e.g., Bycio et al., 1987) and that supervisors are motivated to complete it accurately. However, this may not always be the case and supervisors may view formal performance appraisals as more of an administrative burden than a useful tool for developing subordinates (e.g., Meyer, 1991), and thus do whatever they can to simply the task for themselves. By forcing raters to recall behaviors to justify a distribution of target performance these "shortcuts" and the associated errors may be minimized. Moreover, the participants in this study received training prior to completing the rating task; something that not all real world supervisors receive prior to evaluating subordinate performance.

In addition, the inclusion of rater training added to the completion time of this study. It is possible that, although prepared to provide more accurate ratings as the result of the training, participants became bored with the task and did not necessarily use those skills. Again, it is unclear whether these issues are unique to undergraduate samples, or if supervisors also experience these same issues in real world situations when, for example, they are required to evaluate the performance of a large number of subordinates in a relatively short period of time.

Future Directions

Future research should continue to focus on true score accuracy as a criterion for evaluating rating formats, including DA. In addition, research should investigate which is the best way to create comparable true scores across DA and TA conditions. Efforts were made to create comparable true scores in the current study, but it may be possible that more appropriate true score development processes exist.

A major question that remains is whether DA will show any practical differences in true score accuracy outside of the laboratory. Field-based findings regarding DA have been mostly positive (i.e., Deadrick & Gardner, 1997; Fox et al., 2005). However, it remains to be seen whether or not DA has an advantage over TA in terms of true score accuracy in actual supervisors. This study included a short rating training exercise prior to the rating task. However, the training did not focus specifically on effectively using either of the two response formats, nor did it allow for participants to practice using the response format and receive feedback on rating performance. Given that most of the DA users were likely much less familiar with the DA response format than their TA counterparts were using the TA response format, more specific rater training may lead to more accurate performance rating. Thus, I suggest that future research on DA provide DA users with more rigorous training including practice and feedback, which will help users familiarize themselves with the format.

Another future direction that should be explored is the practical implications of the additional performance variability information included in DA ratings. Kane (2000) suggested several potential advantages associated with the use of DA: (list them here). However, past research has focused exclusively on rater accuracy (e.g., Deadrick & Gardener, 1997; Fox et al., 2005). This study was the first to begin exploring other potential advantages of DA (i.e., cognitive load – not supported), but no research has yet examined the potential value of the additional information provided through DA ratings. I suggest that this information may prove to be valuable for developing employees by highlighting inconsistent performance and helping employees and their supervisors identify under what conditions or in what situations employees perform particularly well

or not. This information may then be used to improve on specific aspects of performance.

With regard to the findings about cognitive load and the mediational hypothesis, future research should examine participants' cognitive processing ability. As mentioned above, it may be that DA is simply too difficult for some to effectively use. It may prove to be a great advantage for those with the cognitive processing ability to manage DA's cognitive load requirements.

Conclusion

This study extends past research on the advantages of DA for interrater agreement to include true score accuracy and the use of trained raters. Although findings suggest that DA is advantageous overall, findings from the current study do not align with previous research regarding the advantages of DA in situations of inconsistent performance (e.g., Deadrick & Gardner, 1997; Steiner et al., 1993). In fact, findings using interrater agreement and true score accuracy as criteria suggest that TA may have a significant advantage over DA. Second, past research has suggested that DA users experience less cognitive load than TA users. This study was the first to empirically test this proposition, and findings strongly suggest that DA users experience significantly greater, and not less, cognitive load during the rating process. Finally, it appears that cognitive load does not mediate the relationship between response format type and rater accuracy. Thus, future research may aim to identify under what conditions DA is advantageous over TA, and why DA users experience more cognitive load during the rating process. Understanding these issues may benefit performance rating research as a whole.

REFERENCES

Atkinson, R. C., & Shiffin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation: advances in research and theory* (Vol. 2). New York: Academic Press.

Austin, J. T., & Villanova, P. (1992). The criterion problem: 1917-1992. *Journal of Applied Psychology, 77,* 836-874.

Ayres, P. (2006). Using subjective measures to detect variations of intrinsic cognitive load within problems. *Learning and Instruction, 16,* 389-400.

Barrett, R. S., Taylor, E. K., Parker, J. W., & Martens, L. (1958). Rating scale content: I. Scale information and supervisory ratings. *Personnel Psychology, 11,* 333-346.

Battiste, V., & Bortolussi, M. (Eds.). (1988). Transport pilot workload: a comparison of two subjective techniques. Proceedings from *The Human Factors Society Thirty-second Annual Meeting* (pp. 150-154). Santa Monica, CA: Human Factors Society.

Beal, D. J., Weiss, H. M., Barros, E., & MacDermid, S. M. (2005). An episodic process model of affective influences on performance. *Journal of Applied Psychology, 90,* 1054-1068.

Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review, 6,* 205-212.

Bittner, A. V., Byers, J. C., Hill, S. G., Zacklad, A. L., & Christ, R. E. (1989). Generic workload ratings of a mobile air defence system (LOS-F-H). In Proceedings from *The Human Factors Society Thirty-third Annual Meeting* (pp. 1476-1480). Santa Monica, CA: Human Factors Society.

Borman, W. C. (1975). Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. *Journal of Applied Psychology, 60,* 556-560.

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

Borman, W. C. (1979). Format and training effects on rating accuracy and rating errors. *Journal of Applied Psychology, 64,* 412-421.

Bycio, P., Alvares, K. M., & Hahn, J. (1987). Situational specificity in assessment center ratings: a confirmatory factor analysis. *Journal of Applied Psychology, 72*, 463-474.

Cao, A., Chintamani, K. K., Pandya, A. K., & Ellis, R. D. (2009). NASA TLX: software for assessing subjective mental workload. *Behavior Research Methods, 41*, 113-117.

Cardy, R. L., & Dobbins, G. H. (1994). *Performance appraisal: Alternative perspectives*. Cincinnati, OH: Southwest.

Cardy, R. L., & Keefe, T. J. (1994). Observational purpose and evaluative articulation in frame-of-reference training: the effects of alternative processing modes on rating accuracy. *Organizational Behavior and Human Decision Processes, 57*, 338-357.

Corwin, W. H., Sandry-Garza, D. L., Biferno, M. H., Boucek, G. P., Logan, A. L., Jonsson, J. E., & Metalis, S. A. (1989). Assessment of crew workload measurement methods, techniques, and procedures: Process methods and results. *Report WRDC-TR-89-7006*. Wright-Patterson Air Force Base, OH: Wright Research and Development Center, Air Force Systems Command.

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition, 58*, 1-73.

Cronbach, L. J. (1955). Processes affecting scores on "understanding of others" and "assumed similarity". *Psychological Bulletin, 52*, 177-193.

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*, 158-167.

Deadrick, D. L., & Gardner, D. G. (1997). Distributional ratings of performance levels and variability: an examination of rating validity in a field setting. *Group and Organization Management, 22*, 317-342.

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*, 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*, 360-396.

DeNisi, A. S., & Williams, K. J. (1988). Cognitive approaches to performance appraisal. In G. Ferris & K. Rowland (Eds.), *Research in personnel and human resource management* (Vol. 6). Greenwich, CT: JAI Press.

DuBois, C. L. Z., Sackett, P. R., Zedeck, S., & Fogli, L. (1993). Further exploration of typical and maximum performance criteria: definitional issues, prediction, and white-black differences. *Journal of Applied Psychology, 78*, 205-211.

Edwards, B. D., & Woehr, D. J. (2007). An examination and evaluation of frequency-based personality measurement. *Personality and Individual Differences, 43*, 803-814.

Eggemeier, F. T., & Wilson, G. F. (1991). Performance-based and subjective assessment of workload in multi-task environments. In D. L. Damos (Ed.), *Multiple task performance* (pp. 217-278). London: Taylor & Francis Group.

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

Fox, S., Bizman, A., & Garti, A. (2005). Is distributional appraisal more effective than the traditional performance appraisal method? *European Journal of Psychological Assessment, 21*, 165-172.

Fritz, M. S., & McKinnon, D. P. (2007). Required sample size to detect the mediated effect. *Psychological Science, 18*, 233-239.

Gaugler, B. B., & Thornton, G. C., III (1989). Number of assessment center dimensions as a determinant of assessor accuracy. *Journal of Applied Psychology, 74*, 611-618.

Gorman, C. A., & Rentsch, J. R. (2009). Evaluating frame-of-reference rater training effectiveness using performance schema accuracy. *Journal of Applied Psychology, 94*, 1336-1344.

Guion, R. M. (1965). *Personnel testing*. New York: McGraw-Hill.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati (Eds.), *Human mental workload* (pp. 139-183). Amsterdam: North Holland Press.

Hasher, L., & Zacks, R. T. (1984). Automatic processing of fundamental information: the case of frequency of occurrence. *American Psychologist, 39*, 1372-1388.

Hastie, R., & Park, B. (1986). The relationship between memory and judgment depends on whether the judgment task is memory-based or on-line. *Psychological Review, 93*, 258-268.

Heilman, M. E., & Saruwatari, L. R. (1979). When beauty is beastly: the effects of appearance and sex on evaluations of job applicants for managerial and nonmanagerial jobs. *Organizational Behavior and Human Performance, 23*, 360-372.

Hennessy, J., Mabey, B., & Warr, P. (1998). Assessment centre observation procedures: an experimental comparison of traditional, checklist and coding methods. International *Journal of Selection and Assessment, 6*, 222-231.

Ilgen, D. R., & Feldman, J. M. (1983). Performance appraisal: A process focus. In L. Cummings & B. Staw (Eds.), *Research in organizational behavior* (Vol. 5). Greenwich, CT: JAI Press.

Jako, R. A., & Murphy, K. R. (1990). Distributional ratings, judgment decomposition, and their impact on interrater agreement and rating accuracy. *Journal of Applied Psychology, 75*, 500-505.

Judge, T. A., & Ferris, G. R. (1993). Social context of performance evaluation decisions. *Academy of Management Journal, 36*, 80-105.

Kane, J. S. (1986). Performance distribution assessment. In R. A. Berk (Ed), *Performance Assessment: Methods and Applications* (pp. 237-273). Baltimore: Johns Hopkins University Press.

Kane, J. S. (2000). Accuracy and its determinants in distributional assessment. *Human Performance, 13*, 47-84.

Kane, J. S., & Woehr, D. J. (2006). Performance measurement reconsidered: An examination of frequency estimation as a basis for assessment. In W. Bennett, C. Lance, & D. J. Woehr (Eds.), Pe*rformance measurement: Current perspectives and future challenges* (pp. 77-110). Hillsdale, NJ: Lawrence Erlbaum.

Kingsbury, F. A. (1922). Analyzing ratings and training raters. *Journal of Personnel Research, 1*, 377-382.

Kolk, N. J., Born, M. P., van der Flier, H., & Olman, J. M. (2002). Assessment center procedures: cognitve load during the observation phase. *International Journal of Selection and Assessment, 10*, 271-278.

Lado, A. A., & Wilson, M. C. (1994). Human resource systems and sustained competitive advantage: A competency-based perspective. *Academy of Management Review, 19*, 699-727.

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

Landy, F. J., & Rastegary, H. (1988). Current issues in performance evaluation. In I. Robertson & M. Smith (Eds.). *Personnel evaluation of the future*. New York: Wiley.

Latham, G. P., & Wexley, K. N. (1977). Behavioral observation scales for performance appraisal purposes. *Personnel Psychology, 30*, 255-268.

Latham, G. P., Wexley, K. N., & Pursell, E. D. (1975). Training managers to minimize rating errors in the observation of behavior. *Journal of Applied Psychology, 60*, 550-555.

LeBreton, J. M., Wu, J., & Bing, M. N. (2008). The truth(s) on testing for mediation in the social and organizational sciences. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences* (pp. 107-141). New York: Taylor & Francis Group.

Leventhal, L., Turcotte, S. J. C., Abrami, P. C., & Perry, R. P. (1983). Primacy/recency effects in student ratings of instruction: a reinterpretation of gain-loss effects. *Journal of Educational Psychology, 75*, 692-704.

Lord, R. G. (1985). An information processing approach to social perceptions, leadership perceptions and behavioral measurement in organizational settings. In B. M. Staw & L. L. Cummings (Eds.), *Research in organizational behavior* (Vol. 7). Greenwich, CT: JAI Press.

MacDonald, H. A., & Sulsky, L. M. (2009). Rating formats and rater training redux: a context-specific approach for enhancing the effectiveness of performance management. *Canadian Journal of Behavioural Science, 41*, 227-240.

Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: thinking categorically about others. *Annual Review of Psychology, 51*, 93-120.

Madden, J. M., & Bourdon, R. D. (1964). Effects of variations in rating scale format on judgment. *Journal of Applied Psychology, 48*, 147-151.

McArthur, L. (1980). What grabs you? The role of attention in impression formation causal attribution. In E. Higgins, C. Herman, & M. Zanna (Eds.), *Social cognition: The Ontario symposium on personality and social psychology* (Vol. 3). Hillsdale, NJ: Lawrence Erlbaum.

McNamara, T. P. (1999). Single-code versus multiple-code theories in cognition. In R. J. Sternberg (Ed.), *The nature of cognition* (pp. 113-135). Cambridge, MA: MIT Press.

Meyer, H. H. (1991). A solution to the performance appraisal feedback enigma. *Academy of Management Executive, 5*, 68-76.

Murphy, K. R., Balzer, W. K., Lockhart, M., & Eisenman, E. (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.

Murphy, K. R., Gannett, B. A., Herr, B. M., & Chen, J. A. (1986). Effects of subsequent performance on evaluations of previous performance. *Journal of Applied Psychology, 71*, 427-431.

Murphy, K. R., & Pardaffy, V. A. (1989). Bias in behaviorally anchored rating scales: global or scale-specific. *Journal of Applied Psychology, 74*, 343-346.

Nataupsky, M., & Abbott, T. S. (1987). Comparison of workload measures on computer-generated primary flight displays. Proceedings from *The Human Factors Society Thirty-first Annual Meeting* (pp. 548-552). Santa Monica, CA: Human Factors Society.

Nisbett, R., & Ross, L. (1980). *Human inference strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice Hall.

Paas, F. G. W. C. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive-load approach. *Journal of Educational Psychology, 84*, 429-434.

Paas, F. G. W. C., & Van Merrienboer, J. J. G. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review, 6*, 351-371.

Padgett, M. Y., & Ilgen, D. R. (1989). The impact of ratee performance characteristics on rater cognitive processes and alternative measures of rater accuracy. *Organizational Behavior and Human Decision Processes, 44*, 232-260.

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. Psychological Methods, 15, 209-233.

Pulakos, E. D. (1984). A comparison of rater training programs: error training and accuracy training. *Journal of Applied Psychology, 69*, 581-588.

Pursell, E. D., Dossett, D. L., & Latham, G. P. (1980). Obtaining valid predictors by minimizing rating errors in the criterion. *Personnel Psychology, 33*, 91-96.

R Development Core Team. (2005). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Reb, J., & Cropanzano, R. (2007). Evaluating dynamic performance: the influence of salient gestalt characteristics on performance ratings. *Journal of Applied Psychology, 92*, 490-499.

Reilly, R. R., Henry, S., & Smither, J. W. (1990). An examination of the effects of using behavior checklists on the construct validity of assessment center dimensions. *Personnel Psychology, 43*, 71-84.

Ruchkin, D. S., Grafman, J., Cameron, K., & Berndt, R. S. (2003). Working memory retention systems: a state of activated long-term memory. *Behavioral and Brain Sciences, 26*, 709-777.

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

Sawin, D. A., & Scerbo, M. W. (1995). Effects of instruction type and boredom proneness in vigilance: implications for boredom and workload. *Human Factors, 37*, 752-765.

Schleicher, D. J., Day, D. V., Mayes, B. T., & Riggio, R. E. (2002). A new frame for frame-of-reference training: enhancing the construct validity of assessment centers. *Journal of Applied Psychology, 87*, 735-746.

Sears, D. O. (1986). College sophomores in the laboratory: influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology, 51*, 515-530.

Sedlmeier, P., Hertwig, R., & Gigerenzer, G. (1998). Are judgments of the positional frequencies of letters systematically biased due to availability. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 754-770.

Shively, R., Battiste, V., Matsumoto, J., Pepiton, D., Bortolussi, M., & Hart, S. G. (1987). In flight evaluation of pilot workload measures for rotorcraft research. Proceedings from *The Fourth Symposium on Aviation Psychology* (pp. 637-643). Columbus, OH: Department of Aviation, Ohio State University.

Simonson, I., & Nye, P. (1992). The effect of accountability on susceptibility to decision errors. *Organizational Behavior and Human Decision Processes, 51*, 416-446.

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: an approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology, 47*, 149-155.

Steiner, D. D., Rain, J. S., & Smalley, M. M. (1993). Distributional ratings of performance: further examinations of a new rating format. *Journal of Applied Psychology, 78*, 438-442

Sweller, J. (2006). Natural information processing systems. *Evolutionary Psychology, 4*, 434-458.

Task Force. (2009). Guidelines and ethical considerations for assessment center operations. *International Journal of Selection and Assessment, 17*, 243-253.

Taylor, S. E., & Fiske, S. T. (1978). Salience, attention, and attributions: top of the head phenomena. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 11). New York: Academic Press.

Taylor, E. K., & Wherry, R. J. (1951). A study of leniency in two rating systems. *Personnel Psychology, 4*, 39-47.

Thornton, G. C., III, & Rupp, D. E. (2005). *Assessment centers in human resource management*. Mahwah, NJ: Erlbaum.

Tsang, P. S., & Johnson, W. W. (1989). Cognitive demand in automation. *Aviation, Space, and Environmental Medicine, 60*, 130-135.

Tulving, E. (1983). *Essentials of episodic memory*. New York: Oxford University Press.

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

Vidulich, M. A., & Bortolussi, M. R. (1988). A dissociation of objective and subjective workload measures in assessing the impact of speech controls in advanced helicopters. Proceedings from *The Human Factors Society Thirty-second Annual Meeting* (pp. 1471-1475). Santa Monica, CA: Human Factors Society.

Wagner, S., & Goffin, R. (1997). Differences in accuracy of absolute and comparative performance appraisal methods. *Organizational Behavior and Human Decision Processes, 70*, 95-103.

Wells, F. L. (1907). A statistical study of literary merit. *Archives of Psychology, 7*.

Woehr, D. J. (1992). Performance dimension accessibility: implications for rating accuracy. *Journal of Organizational Behavior, 13*, 357-367.

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: a quantitative review. *Journal of Occupational and Organizational Psychology, 67*, 189-205.

Woehr, D. J., & Miller, M. J. (1997). Distributional ratings of performance: more evidence for a new rating format. *Journal of Management, 5*, 705-720.

Zedeck, S. (1986). A process analysis of the assessment center method. In B. M. Staw & L. L. Cummings (Eds.), *Research in organizational behavior, Vol. 8* (pp. 259-296). Greenwich, CT: JAI Press.

APPENDIX A.

DA response format for the skill dimension of communication

Instructions: Respond to the following questions by marking the percentage of time the target spends displaying the behaviors described below. Remember that combining all four categories must add up to 100%.

4	3	2	1
____% of time target displays the following behaviors	____% of time target displays the following behaviors	____% of time target displays the following behaviors	____% of time target displays the following behaviors
Speaks clearly and fluently; Consistently makes direct eye contact, and nonverbal behavior is open and friendly; Communication is polite and respectful; Avoids jargon.	Usually speaks clearly and fluently; Sometimes makes eye contact; nonverbal behavior is neutral; tries to explain own ideas, but may be vague or overly wordy at times; is polite.	Difficult to understand because of speed or volume of speech; little eye contact; fidgets or uses other distracting nonverbal behavior; some sentences are convoluted or incomplete; misuses words.	Nonverbal behavior indicates boredom or hostility; sentences are unconnected or incomplete; difficult to understand meaning of statements; frequently misuses words; is harsh, rude, or overly personal with teammates.

APPENDIX B

TA response format for the skill dimension of communication

Instructions: Respond to the following questions by marking the number that best represents the behaviors described below. Remember to only mark a single number.

4	3	2	1
_____	_____	_____	_____
Speaks clearly and fluently; Consistently makes direct eye contact, and nonverbal behavior is open and friendly; Communication is polite and respectful; Avoids jargon.	Usually speaks clearly and fluently; Sometimes makes eye contact; nonverbal behavior is neutral; tries to explain own ideas, but may be vague or overly wordy at times; is polite.	Difficult to understand because of speed or volume of speech; little eye contact; fidgets or uses other distracting nonverbal behavior; some sentences are convoluted or incomplete; misuses words.	Nonverbal behavior indicates boredom or hostility; sentences are unconnected or incomplete; difficult to understand meaning of statements; frequently misuses words; is harsh, rude, or overly personal with teammates.

APPENDIX C

Sample rating target script for AC simulation exercise

This script is for the high and consistently performing rating target

-be the first to introduce yourself and initiate group discussion

-actively make eye contact when initiating group discussion

-offer to take notes and actively make eye contact when others provide input

-present three different useful ideas during group discussion, bring up:
    -that each dorm should have their own colored headbands or t-shirts
    -that KCSU DJs the event
    -that food (e.g., pizza, tacos) are provided

-When others bring up ideas be sure to let them know their ideas are valued (do this to varying degrees), e.g.,:
    -“yeah that is a really good idea [elaborate on idea]”
    -[nod and smile]
    -“good idea!”
    -in general use responses that range from “very positive” to “mildly positive”

APPENDIX D

Conscientiousness scale in Big Five Inventory.

(5-point scale: 1=disagree strongly and 5=agree strongly).
I see myself as…
1. someone who does a thorough job.
2. Can be somewhat careless.
3. Is a reliable worker.
4. Tends to be disorganized.
5. Tends to be lazy.
6. Perseveres until the task is finished.
7. Does thing efficiently.
8. Makes plans and follows through with them.
9. Is easily distracted.