5-2012

The Relationship Between Rater Agreement, Behavioral Observability and Overall Impressions

Jennifer N. Scott
Western Kentucky University, jennifer.scott223@topper.wku.edu

THE RELATIONSHIP BETWEEN RATER AGREEMENT, BEHAVIORAL OBSERVABILITY AND OVERALL IMPRESSIONS

A Thesis
Presented to
The Faculty of the Department of Psychology
Western Kentucky University
Bowling Green, Kentucky

In Partial Fulfillment
Of the Requirements for the Degree
Master of Arts

By
Jennifer N. Scott

May 2012

THE RELATIONSHIP BETWEEN RATER AGREEMENT, BEHAVIORAL OBSERVABILITY AND OVERALL IMPRESSIONS

Date Recommended 5/2/12

Anthony R. Paquin, Director of Thesis

Reagan Brown

Aaron Wichman

Dean, Graduate Studies and Research 18-May-2012 Date

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Anthony R. Paquin, for his help and support with my thesis project. I would not have been able to complete it without him. Also, I would like to thank my committee members Reagan Brown and Aaron Wichman for their help and assistance with this project. They both provided valuable insights and contributed to the success of my thesis.

CONTENTS

THE RELATIONSHIP BETWEEN RATER AGREEMENT, BEHAVIORAL OBSERVABILITY AND OVERALL IMPRESSIONS

Jennifer Scott May 2012 22 Pages

Directed by Anthony R. Paquin, Reagan Brown, and Aaron Wichman

Department of Psychology Western Kentucky University

This study examined two item characteristics believed to influence rater agreement: observability and difficulty. The first goal of this study was to replicate the findings of Roch, Paquin and Littlejohn (2009), which found that rater agreement was negatively related to item observability (Hypothesis 1) and rating difficulty (Hypothesis 2). The study also explored whether participants had closer item performance ratings to their overall impression when items were less observable (Hypothesis 3) and more difficult to rate (Hypothesis 4). A sample of 254 Undergraduate psychology students viewed a video of a leaderless group discussion and then filled out a rating form assessing performance of one of the individuals in the video and rating difficulty. Results were that rater agreement was positively related to observability (not supporting Hypothesis 1) and negatively related to difficulty (supporting Hypothesis 2). RDS, a distance score between participant's overall impression and the item performance rating was computed to assess Hypotheses 3 and 4. RDS was positively related to observability (supporting Hypothesis 3) and not related to difficulty (not supporting Hypothesis 4). The positive relationship between observability and rater agreement was surprising given that it was the opposite of previous findings. Not hypothesized but of interest to the study was that observability and difficulty were not correlated. In previous studies, these variables were negatively correlated. Implications of these findings are discussed along with directions for further research.

Rater Agreement Overview

Using multiple raters is essential in many aspects of industrial/organizational psychology practices, including 360-degree feedback, performance appraisals, interviews, assessment centers, developing behaviorally anchored rating scales, and determining critical job dimensions (Kozlowski & Hattrup, 1992; Roch, Paquin and Littlejohn, 2009). The idea in these practices and others using multiple raters is that having ratings from only one source may not be as good as having ratings from multiple sources for the same object. Multiple ratings can increase the amount of information received because different people may have different perspectives on behaviors or different opportunities to observe individuals. In addition, there are situations when having higher agreement between raters would be important, such as in selection procedures.

Although interrater agreement is a somewhat widely studied area in industrial/organization psychology, measuring rater agreement has been somewhat problematic because the term has often been erroneously used to refer to interrater reliability. Interrater agreement is the extent to which different raters will give the same scores to an individual being rated. For instance, assume two supervisors are rating the same employee for her annual performance evaluation, and the rating form consists of five items with a Likert-type scale ranging from 1 to 5. If Supervisor 1 gives the employee scores of 2, 3, 4, 3, and 3, to have perfect agreement, Supervisor 2 would also need to give the employee scores of 2, 3, 4, 3 and 3. Interrater reliability, however, involves the pattern of scores raters give. It essentially uses correlations between the two (or more) raters' scores. In the example above, if Supervisor 1 gives the same scores,

Supervisor 2 could give scores of 3, 4, 5, 4, and 4 and have perfect interrater reliability since the pattern of scores is the same. The two supervisors, however, would have no interrater agreement because they did not provide the same score for any of the items.

Why Study Rater Agreement?

It is important to study rater agreement because many business processes use multiple raters. These situations may benefit from multiple raters because more raters potentially provide a more complete view of the person being rated. Other times, using multiple raters is useful in gauging the accuracy of the ratings when characteristics of the ratee are known. Understanding the different reasons raters disagree can help make those processes more accurate.

The conventional idea is that higher agreement between raters is better. Higher agreement supposedly indicates the presence of less measurement error and better quality in the ratings (e.g. Feldman, 1981; Roch et al., 2009). The question then becomes whether or not rater disagreement actually equals error in measurement or something else. This has concerned many psychologists recently, and the conclusion that the research seems to suggest is that there are many reasons raters disagree.

One reason raters may disagree is because of differences in the characteristics of the raters (e.g. Feldman, 1981; Murphy, Cleveland, Skattebo, & Kinney, 2004). People are all unique, so it makes sense that their differences could also influence the rating process. There are various characteristics of raters that can cause interrater disagreements. One of these is the rater's cognitive processes (Feldman, 1981).

Cognitive Processes

Borman (1978) suggested that there are three steps involved in making performance evaluations: observing work-related behavior, evaluating these behaviors, and weighting the evaluations to arrive at a single rating on a performance dimension. Rater disagreement, he suggested, can spring from differences in any of these steps. For instance in the first step, observing employee behavior, raters may have different opportunities to observe behavior, or, given the same opportunities, they may attend to different aspects of that behavior. When evaluating behaviors, raters can differ in the importance they place on specific behaviors and the effectiveness of those behaviors (Borman, 1978). According to Borman (1978), raters may use previous experience to form opinions about behaviors, or use a global opinion to influence their beliefs about individual ratings. Differences also occur when raters have to make judgments regarding individual behaviors because raters may weigh some performances more heavily or rely on first impressions (Borman, 1978).

Along the same line as Borman (1978), Feldman (1981) suggested that the cognitive processes raters go through may influence individual differences in ratings, and thus rater agreement. Feldman (1981) also noted that there are stages a rater must go through to rate an employee. The rater must first recognize and attend to employee behaviors, then organize and store the information, and later recall relevant information when judgments are required. Supervisors, in addition, must integrate all the information about the ratee they have into a summary judgment. This can be made at any stage, and judgments could change throughout the process. (Feldman, 1981). For instance, a supervisor may recognize an employee behaving in a certain way and immediately make

a judgment about the behavior. This judgment is then stored with the behavior in the supervisor's mind, and both the behavior and judgment are recalled when the supervisor has to make performance ratings.

After noticing behaviors and forming judgments, raters must recall the information to make performance ratings. Feldman (1981) stated that when raters give performance ratings, they rely on categorization. One interesting aspect of categorization is using a prototype, or typical example of the category. Feldman noted that when no information is available, the rater might guess about the characteristics of the ratee using the prototype for the category in which they believe the ratee is located. It may not be that the rater is consciously aware of guessing, and it is possible that they have formed false memories that are consistent with their beliefs about the people in the category with the ratee (Feldman, 1981).

Relying on categorization and prototypes of groups when sufficient information is not available about ratees has implications for rater agreement. These implications stem from differences and similarities in the raters. If the raters use different categories to assess ratees, rater agreement might suffer (Feldman, 1981). If, on the other hand, multiple raters use the same category and exemplar of that category as the basis for their ratings, then rater agreement could be high (Feldman, 1981).

Along the same lines as the categorization process raters go through, is a documented type of rater error called halo error. This type of error was one of the first studied and still is one of the most often mentioned types of rater error. Halo error was originally defined as when ratings of individual items are influenced by the rater's overall feelings toward the person being rated (Thorndike, 1920). This means that raters who

like the person being rated could rate him higher than he deserves, if the raters succumb to halo error. Since Thorndike's introduction of halo error, many researchers have expanded on and revised the definition in an attempt to understand halo error better and evaluate the effects it has on performance ratings (Balzer & Sulsky, 1992). Jacobs and Kozlowski (1985) identified four different methods of analyzing halo error in research literature but found that all the methods were equally consistent over time.

Halo error is important to understand because it provides an example of how people use overall judgments to make ratings on specific items. Some researchers looking at rater error have proposed that correcting halo error would improve the quality of ratings (Balzer & Sulsky, 1992). This would imply that improving the quality of ratings would reduce interrater disagreement.

Item Characteristics

Not only can rater characteristics impact rater agreement, but characteristics of the items on a rating form can contribute to rating discrepancies (e.g. Kaiser & Craig, 2005; Roch et al., 2009). Indeed, Kaiser and Craig (2005) proposed that a possible reason for discrepancies among raters could be the items used in performance appraisals. Items might be written in a way that invites different interpretations, and researchers are only beginning to investigate this idea. Towards this end, a number of researchers have recently been interested in understanding different item characteristics that could influence rater agreement (e.g., Roch et al., 2009). Among these characteristics are syntax, multibarreledness, behavioral specificity, perceived difficulty and behavioral observability (Brutus & Facteau, 2003; Kaiser & Craig, 2005; Roch et al. 2009).

Syntax, Multibarreledness and Behavioral Specificity

Brutus and Facteau (2003) introduced the idea that an item's syntax, multibarreledness and behavioral specificity could influence the psychometric quality of the item. Psychometric quality was defined as the relationship between the item and the performance dimension it is supposed to measure as determined by confirmatory factor analysis. Syntax was characterized by the number of basic sentence parts, or constituents, an item has, and how these constituents are ordered and grouped together. An item was multi-barreled if it had more than one behavioral referent in the item, and behavioral specificity was measured by the degree to which the item "narrowly identifies the behavior to be evaluated [and] also provides, when possible, a contextual frame within which the target behavior is expected to occur" (p.315). The results of the study were that syntax was partially related to psychometric quality, but behavioral specificity and multibarreledness were not.

Kaiser and Craig (2005) built upon Brutus and Facteau's (2003) study and looked specifically at syntax, multibarreledness and behavioral specificity, as defined in Kaiser and Craig's study, in relation to interrater reliability and agreement. Kaiser and Craig hypothesized that number of constituents (basic sentence parts), behavioral references and the degree of behavioral abstraction would each be negatively related interrater agreement. They tested their hypothesis using a sample of managers who participated in optional leadership development courses. Each of the 1,404 target managers had two superior ratings, two peer ratings and two subordinate ratings on a 360° feedback instrument. All of the items on the instrument were assessed by experts with respect to their syntax, multibarreledness and behavioral specificity. Results indicated that interrater

reliability showed a significant negative correlation with multibarreledness, and nearly significant correlations with the other two item characteristics. None of the three item characteristics were significantly related to interrater agreement (Kaiser & Craig, 2005). The researchers, however, stated that their results could be attributed to low variability in the measures of syntax, multibarreledness and behavioral specificity. Another possible explanation for the nonsignificant results was multi-dimensionality of the items contained used in the study (Roch et al., 2009).

Difficulty

Another characteristic that could affect rater agreement is the item's perceived rating difficulty. Wohlers and London (1989) examined the extent to which the perceived rating difficulty of an item would influence rater agreement. In their first study, they had raters rate the difficulty of items on 30 dimensions of a performance rating scale. The items that were seen as more difficult to rate were also from categories that had less observable behaviors. The difficulty ratings were used for the second study in which Wohlers and London hypothesized that there would be lower agreement on managerial characteristics that were perceived as more difficult to rate. Their sample included 52 middle-level managers who were each asked to hand out the performance questionnaire to three each of lower level, higher level and same level coworkers. The results supported the hypothesis. Items that were more difficult to rate had higher levels of variability between ratings and thus lower interrater agreement (Wohlers & London, 1989). Wohlers and London (1989) also noted that there was a great deal of variability (more disagreement among raters) in some items raters perceived to be easy to rate. They proposed that raters might disagree on the items perceived as easy to rate because they

use different definitions of a term but believe their definition to be correct. This would cause the perception of an easy rating, but would also show high variability in ratings.

Another study looking at difficulty of items and interrater agreement was conducted by Roch et al. (2009). They used undergraduate students who first watched a video of an assessment center role-playing activity and then rated the target individual. The difficulty of items on the rating form was assessed by a number of experts in assessor training, and each item was given a difficulty score. They found that raters actually agreed more on items that were more difficult.

Behavioral Observability

Roch et al. (2009) also examined the behavioral observability of the items on the rating form. They argued that the concept of behavioral specificity used in Kaiser and Craig's (2005) study actually contained three dimensions: whether the behavior is observable, whether it is narrowly defined, and whether context is provided. Roch et al. decided to examine just the extent to which a behavior is observable. They stated that an item could be behaviorally specific but have low behavioral observability, hence the need to study the concepts separately. As with perceived difficulty, experts rated the behavioral observability of the items using a scale that ranged from observable behaviors to subjective judgment, and an observability score was achieved for each item.

The results of Roch et al.'s (2009) studies were that as behavioral observability increased, rater agreement decreased, and as difficulty increased, rater agreement increased. They also found that more observable behaviors were seen as easier to rate. Roch et al. suggested that raters might be defaulting to their overall impression of the target when items become difficult to rate and less observable, which, if the overall

impressions of the raters are similar, could increase the likelihood that raters would agree on these items. Unfortunately, as the researchers did not measure overall impressions, this speculation could not be tested.

Can different item characteristics such as difficulty and observability influence raters' ratings and cause them to use an overall evaluation of the ratee when rating specific items? As has been previously stated, item difficulty and behavioral observability can influence rater agreement (Brutus & Facteau, 2003; Kaiser & Craig, 2005; Roch et al., 2009; Wohlers & London, 1989), but there is little research on how those specific item characteristics could have an effect on people reverting to overall judgments.

There is some evidence that a rater's overall judgment could influence their ratings on specific items. Feldman (1981), for example, noted that when raters cannot access specific information about an employee they may use their general feeling of the employee or characteristics of the prototype of the group for which that employee belongs in the rater's mind.

Raters might likely perceive items for which they could not access relevant information regarding the ratee to be more difficult, and thus rely on general feelings or prototypical information. This would lead to raters using prototypes more often for items that are difficult than for items that are easy to rate.

Purpose of This Study

This study will attempt to replicate the findings of Roch et al. (2009) involving rater agreement, item difficulty and behavioral observability. Thus the first two hypotheses will be the same as those used in the Roch et al. study:

1. Rater agreement will be negatively correlated with behavioral observability.

2. Rater agreement will be positively correlated with perceived item difficulty.

If the results indicate that lower behavioral observability and higher difficulty of items are related to more interrater agreement (i.e., replicate Roch et al. 2009), the question becomes why this is the case. From the research above, the argument can be made that a rater's overall judgments may be affecting their ratings for specific items (Balzer & Sulsky, 1992; Feldman, 1981). The last hypotheses, therefore, will address whether raters are defaulting to an overall judgment when items become less observable and more difficult:

3. Items that are less observable will have ratings that are closer to the overall judgment of the person being rated than more observable items
4. Items that are more difficult will have ratings that are closer to the overall judgment of the person being rated than less difficult items.

Method

Participants

Participants were 253 undergraduate psychology students (79 male, 174 female, M_{age} = 19.9, age range: 18-55 years) who received class or extra credit for participating.

Materials

Stimulus Performance.

Performance information was presented via a 20-minute video which depicted a leaderless group discussion role-role playing exercise. This was the same video used in Montgomery's (2010) study. The video included four individuals of which three were male and one was female. One of the male participants represented the target performance.

Rating Form.

The participants filled out a form concerning their observations of the target individual from the video. The form included a list of 85 statements about the behaviors of the target individual. For example, a statement on the form is "knew how to solve problems". Participants rated the individual on each of these items using a Likert-type scale ranging from 1 (*Not at all)* to 5 (*To a very great extent).* Participants also reported their difficulty in rating the target individual for each item. This rating also used a Likert-type scale which ranged from 1 (*Very easy to rate*) to 5 (*Very difficult to rate*).

The form was similar to the one used in the study by Roch et al. (2009). This rating form differed from the Roch et al. (2009) form in that the question order was changed and two versions were used to control for effects of question order. To develop two equal sections of the form, all of the questions were ordered according to their observability rating and divided into two sections (Section 1 and Section 2). The sections were created by putting the item with the highest observability score in Section 1, then the items with the next two highest observability scores in Section 2, the next two in Section 1, and so on until all of the items were in either Section 1 or Section 2. Each section had the same average amount of observability (3.10 on a 5-point scale) for each group and the individual item observability was equally distributed. The questions for each section were then ordered using a random number sequence. Half of the forms included Section 1 first and Section 2 second (Form 1), and half of the forms included Section 2 first, and Section 1 second (Form 2).

This rating form also differed from the one used in the original study in that it included an extra question that asked participants to relate their overall feeling of the

target individual with the statement “To what extent do you feel like the person did a good job overall?” Participants answered this question using a Likert-type scale ranging from 1 (*Not at all)* to 5 (*To a very great extent*). In half of the forms, this item was at the beginning (Overall First), and on half of the forms, this question was at the end (Overall Last). These distinctions created four separate versions of the rating form:

1. Form 1; Overall First
2. Form 1; Overall Last
3. Form 2; Overall First
4. Form 2; Overall Last

Each rating form also included some demographic questions after these sections.

Procedure

One to 15 students took part in each experimental session which lasted between 45 minutes and an hour. Participants were given informed consent documents to sign and return before the study began. At the start of the study, participants were informed that they would be watching a video of a leaderless group discussion and that they should pay attention to Individual A. Then they were told that after watching the video they would be given a rating form regarding their observations of the target individual followed by some demographic questions.

Following these instructions, background information on the topic of the video (see Appendix) was read to the participants, and the video was started. After the video, rating forms were distributed to participants. Prior to the session, the rating forms were ordered sequentially. When the rating forms were handed out, the first person received Version 1, the second person received Version 2 and so on so that the rating form each

participant received was randomly assigned, and this pattern continued to the following study sessions.

Results

$R_{wg,}$ a measure of variability between raters, was computed to measure rater agreement for each item, using the method described by James, Demaree, & Wolf (1984). The average R_{wg} across items was .52 (*SD* = .14). Average difficulty ratings were also computed for each item (*M* = 1.81, *SD* = .27). Observability was previously assessed by experts (Roch et al., 2009) and ranged from 1 to 5, with 5 indicating more behavioral observability. The mean overall performance rating was 3.92 (*SD* = .76), and the ratings did not differ significantly by whether the overall impression question was at the beginning (*M* = 3.98, *SD* = .80) or end (*M* = 3.88, *SD* = .71; *p* = .310). There were also no differences in the mean agreement or difficulty ratings between participants who received Form 1 and Form 2, or between having the overall impression question at the beginning or end. Those results can be seen in Table 1.

Table 1

Means and Standard Deviations by Form and Overall Question Placement

	Form		Overall Question Placement	
	1	2	Beginning	End
Variable	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
Agreement	0.52 (0.15)	0.51 (0.15)	0.53 (0.16)	0.51 (0.14)
Difficulty	1.83 (0.28)	1.79 (0.27)	1.81 (0.26)	1.81 (0.27)
Overall Impression	3.85 (0.72)	4.00 (0.78)	3.98 (0.80)	3.88 (0.71)

Results were analyzed first for each form (From 1 and Form 2) and overall impression question placement (beginning or end) separately, and then compared. There were no significant differences between the correlations for any of these comparisons so all data were analyzed together. A summary of the results can be seen in Table 2.

Table 2

Means, Standard Deviations, and Correlations for the Study Variables

Variable	Mean	*SD*	1	2	3	4
1. Agreement	.52	.14	-			
2. Difficulty	1.81	.27	0.29**	-		
3. Observability	3.10	1.16	0.49**	-0.09	-	
4. RDS	1.00	.55	.26*	-.16	0.34**	-

Note. RDS = rating difference score, discussed below.
* $p < .05$. ** $p < .01$

Hypothesis 1 and 2 Results

Rater agreement was positively correlated with observability ($r = .49$, $p < .001$) and item difficulty ($r = .29$, $p = .006$), thus not supporting Hypothesis 1, but supporting Hypothesis 2. It should be noted that observability was significantly correlated with rater agreement but opposite the hypothesized direction.

Hypothesis 3 and 4 Results

To assess Hypotheses 3 and 4, a rating difference score (RDS) was computed using the following equation:

RDS = |Overall Impression – Item Performance Rating|.

RDS is the absolute value of the overall impression rating minus the individual item rating. Each person, therefore, had an RDS for each item for which they gave a performance rating. The average RDS across items was 1.00 (*SD* = .55).

RDS was significantly correlated with observability (r = .336, p = .002), meaning that participants had individual item ratings that were closer to their overall impressions when items were less observable. RDS was not significantly correlated with difficulty (r = -.159, p = .148). Thus, Hypotheses 3 was supported while Hypothesis 4 was not. Also, data was analyzed to see if there was an interaction between RDS, rater agreement and observability, but there was none. While examining this relationship, however, it was discovered that there was a significant nonlinear relationship between rater agreement and RDS (r = .79; p < .001).

Discussion

The goal of this study was to replicate the findings of Roch et al. (2009) and to test a possible explanation for their findings. Towards this end, the relationships between rater agreement, item observability, perceived item difficulty and overall impression of the ratee were examined. The results as they relate to rater agreement (Hypotheses 1 and 2), then as they relate to individuals' overall impression of the ratee (Hypotheses 3 and 4) will be discussed in the following sections.

Rater Agreement

Hypothesis 1, as items were less observable, rater agreement would increase, was not supported. Instead, the results showed that rater agreement and observability were positively correlated – raters agreed more on items that were more observable. Although this result is different from what was expected, there is some support for the idea that observability and agreement would be positively related. Aiken stated (as cited in Brutus and Facteau, 2003) that "In general, errors in rating are smaller if each characteristic or behavior being rated is described as objectively as possible with reference to some

actually observed activity…" (p. 50). However, the idea has not been supported by research until now and is in contrast to previous findings (Montgomery, 2010; Roch et al., 2009) where the results were the exact opposite. This result is also different from a study with a similar variable, behavioral specificity (Kaiser & Craig, 2005). Those researchers hypothesized that behavioral specificity (of which behavioral observability is a part) would be positively related agreement but did not find any connection between rater agreement and behavioral specificity.

Because the observability of the items was not rated by participants in this study, it is possible that their view of the behavioral observability of the items might have been different from those in the previous studies, which, in turn, impacted the results obtained in this study. It may also be that participants in this study were somehow different from the previous studies. While this is a possibility, the exact nature of these differences is unclear, as all of the previous studies used a sample of college students enrolled in psychology classes. Further, this study used students from the same university as two of the three previous studies (Montgomery, 2010; Roch et al., 2009).

Hypothesis 2, which stated that raters would agree more when items were more difficult to rate was supported. This is in line with previous research. There was a slight difference between difficulty and agreement when participants received the overall impression question first versus when they received the overall impression question at the end, but this difference was not significant.

Overall Impressions

The reason it was thought that participants would agree more on items that were less observable and more difficult to rate was because it was believed that they would

revert to their overall impressions when they could not recall enough information to make a good rating (Hypotheses 3 and 4). Concerning overall impressions and behavioral observability, it is interesting that although participants agreed more on more observable items, less observable items had ratings that were closer to the raters' overall impressions. So, as items were more observable, raters agreed more on those items, and tended to give individual item performance ratings that were less like their overall impression of the person being rated. This might seem to contradict the finding that RDS was positively related to rater agreement – participants agreed more on when individual item ratings were closer to their overall impressions of the ratee. However, raters may have agreed more on items with higher observability while still agreeing on items that were less observable, suggesting that raters did rely on their overall impression to make ratings when items were less observable.

The fourth hypothesis was that participants would default to an overall impression of the person being rated when items became more difficult to rate, making the difference between the item performance rating and their overall impression smaller for more difficult items. However, this hypothesis was not supported. It still is possible that participants will default to their overall impressions when items become difficult to rate, given the cognitive processes raters go though (e.g. Feldman, 1981), but this is not necessarily the case if observability and difficulty are not related.

That there was no relationship between difficulty and overall impressions, further illustrates the distinction between observability and difficulty in this study. Participants did not find less observable items more difficult to rate. This suggests that they were using some other criteria for determining what makes an item difficult. It is possible that

participants felt that less observable items were more difficult to rate, but that other aspects of the items overpowered the observability of the items in determining item difficulty. It should be noted though, that while this study did not find any relationship between observability and difficulty, it is contradictory to previous studies (Montgomery, 2010; Roch et al., 2009; Wholers and London, 1989) where difficulty and observability were negatively correlated. Thus, if the results of this study cannot be replicated, the question still remains as to why participants would agree more on items perceived as more difficult to rate, given that they are not defaulting to an overall impression.

Although, not hypothesized, a curvilinear (i.e., inverted-u) relationship was revealed between agreement and RDS ($r = .79$, $p < .001$). While beyond the scope of this study, it is obvious that further research is needed to explore the implications of this result.

Limitations

One possible limitation of this study was the long questionnaire. Participants were required to answer 85 performance and 85 difficulty items, an overall impression question, and 18 demographic/other items. It is possible that by the end of the questionnaire participants did not put as much effort into giving answers that most closely related to their actual feelings. Going along with the length of the questionnaire is the combined length of the study session which consisted of a 20 minute video on a topic that students probably did not care very much about. This could have caused them to lose interest and not try very hard answering the items on the questionnaire. Also, they did not have much of an incentive to answer as accurately as they could because they received study credit by just attending the study session. An important point here is that while

these are limitations, they do not explain the difference in results from previous studies because one sample in Roch et al. (2009) and the participants in Montgomery's (2010) studies used students from the same university.

While the samples were similar, there were some differences between the current study and previous studies. First, in the Roch et al. (2009) studies a different video was used, but the Montgomery (2010) study used the same video as the current study. Second, the question order in the current study was changed to create two forms. So while the questions were the same, the order of the questions in the two forms of the current study was different from all previous studies. Last, an overall impression question was added.

Future Directions

From the mixed results of this study, it is clear that more research is needed to better understand the relationships between rater agreement, perceived difficulty, behavioral observability and overall impressions. Also, due to the fact that some of the findings in the current study were contradictory to previous findings, this study should be replicated. In addition, there may be more variables that influence how difficult someone thinks an item is to rate, and future studies might examine other characteristics of items besides observability and difficulty. Another route of investigation for future research would be to further examine the relationship between rater agreement and RDS, since it was found to be nonlinear.

APPENDIX

Instructions to Participants

You will be viewing a video of a leaderless group discussion. Please observe the behaviors of the target individual. After the video, you will fill out a rating form regarding your observation of the target individual.

Here is some background information about the video. The people in the video are a team of consultants asked to give recommendation to a client concerning their management problems. The team is to discuss the problem and come to an agreement on the most appropriate solution. The situation is that the personnel and accounting office of a manufacturing company are located on the south side of its factory complex. The offices of the plant manager and production control are located on the north side of the complex. Between these offices lies a major part of the production area. On a regular basis, office employees must walk through the production area for meetings and other work-related reasons. Safety rules require all employees to wear safety hats whenever they enter the production area. It is estimated that 70% of all office supervisors and employees disregard the rule and walk through the assembly area without wearing safety hats. The plant manager wants the team to suggest a motivational or educational technique to increase compliance with this safety rule. In addition, the plant manager seeks their recommendation concerning appropriate disciplinary actions to handle noncompliance.

REFERENCES

Balzer, W. K., & Sulsky, L. M. (1992). Halo and performance appraisal research: A critical examination. *Journal of Applied Psychology, 77*(6), 975-985.

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

Brutus, S., & Facteau, J. (2003). Short, simple, and specific: The influence of item characteristics in multi-source assessment contexts. *International Journal of Selection and Assessment, 11*, 313-325.

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

Jacobs, R., & Kozlowski, S. W. (1985). A closer look at halo error in performance ratings. *Academy of Management Journal, 28*(1), 201-212.

James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology, 69*(1), 85-98.

Kaiser, R. B., & Craig, S. B. (2005). Building a better mouse trap: Item characteristics associated with rating discrepancies in 360 degree feedback. *Consulting Psychology Journal: Practice and Research, 57*(4), 235-245.

Kozlowski, S. W., & Hattrup, K. (1992). A disagreement about within-group agreement: Disentangling issues of consistency versus consensus. *Journal of Applied Psychology, 77*(2), 161-167.

Montgomery, K. E. (2010). *The effects of rater training on the relationship between item observability and rater agreement* (Unpublished master's thesis). Western Kentucky University, Bowling Green, Kentucky.

Murphy, K. R., Cleveland, J. N., Skattebo, A. L., & Kinney, T. B. (2004). Raters who pursue different goals give different ratings. *Journal of Applied Psychology, 89*(1), 158-164.

Roch, S. G., Paquin, A. R., & Littlejohn, T. W. (2009). Do raters agree more on observable items? *Human Performance, 22*(5), 391-409.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4*(1), 25-29.

Wohlers, A. J., & London, M. (1989). Ratings of managerial characteristics: Evaluation difficulty, co-worker agreement, and self-awareness. *Personnel Psychology, 42*, 235-261.