8-7-2013 12:00 AM

Subordinate Ratings of Supervisor Performance: Balancing Accountability and Anonymity

Kevin Doyle
The University of Western Ontario

Supervisor
Dr. Richard Goffin
The University of Western Ontario

Graduate Program in Psychology

A thesis submitted in partial fulfillment of the requirements for the degree in Master of Science

SUBORDINATE RATINGS OF SUPERVISOR PERFORMANCE: BALANCING ACCOUNTABILITY AND ANONYMITY

Thesis format: Monograph

by

Kevin Doyle

Graduate Program in Psychology

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science

The School of Graduate and Postdoctoral Studies
The University of Western Ontario
London, Ontario, Canada

Abstract

Multi-source feedback often includes ratings from one's subordinates; however, there is little research on the accuracy of these ratings. With multi-source feedback systems being used more for administrative decisions there is a precedent to test how accurate subordinate ratings are. The present study distinguishes between two types of accountability; appeasement-accountability and accuracy-accountability in an attempt to increase the accuracy of subordinate ratings of job performance. The result was three experimental conditions. The first was the anonymous condition which is in line with current practice; subordinates are typically granted anonymity when submitting ratings about their supervisor. The second was the appeasement-accountability condition, and the third was the accuracy-accountability condition. One hundred and eight participants rated videos of four different trainers' job performance using behaviorally anchored rating scales (BARS). The dependant variables of interest were Cronbach's (1955) accuracy components of differential accuracy, elevation accuracy, stereotype accuracy, and differential elevation. Significant differences were found between Conditions 1 and 3 for Differential Elevation. Additionally, the comparison between Conditions 2 and 3 was nearing significance for Differential Elevation. These findings are indicative of increased accuracy for administrative decisions for those in the accuracy-accountability condition. Possible explanations for the results found, study limitations, and directions for future research are discussed.

Keywords: Performance management; Performance appraisal; Cronbach components; Feedback; Subordinate Ratings; 360 degree feedback; Multi-source feedback.

Table of Contents

List of Tables

List of Figures

List of Appendices

Performance appraisal is a much-debated topic, and is considered unavoidable by many. It is generally an uncomfortable process for both the person providing the ratings and the person receiving the ratings. In the traditional approach, performance appraisal requires supervisors to rate their employees on a number of scales, followed by a meeting with each employee to review and discuss the ratings provided. In this way, supervisors are held accountable for their ratings simply by meeting face-to-face with their employees. Accountability can be defined as "being answerable for performing up to certain prescribed standards, thereby fulfilling obligations, duties, expectations, and other charges" (Schlenker, Britt, Pennington, Murphy, & Doherty, 1994, p.634). In this sense, should the employees take issue with the ratings received, they are given opportunity to express such concerns. Furthermore, one may suggest that supervisors are motivated by this meeting to engage in more deliberate and more cognitively complex decision making when making their ratings (Tetlock, 1985) and this would presumably increase accuracy. Others have suggested that when someone at a higher level, a supervisor for instance, is rating and is accountable to someone at a lower level, a subordinate, he or she may provide inflated ratings as a way to avoid confrontation (Mero, Guidice, & Brownlee, 2007). Clearly there are mixed views on the accuracy of ratings provided by accountable raters, let alone those who are not held accountable at all.

Accountability in Multi-source Feedback Systems. In multi-source or 360-degree feedback systems, the performance appraisal received may include a supervisory rating, a self-rating, as well as a number of peer and/or subordinate ratings. It has been suggested that ratings from different levels may provide different perspectives on the performance of any given employee, which can help guide the development and

improvement process. The general rule, as stated by Balzer, Greguras, and Raymark (2004), is to include three to five ratings for both peers and subordinates should those sources be used. This provides a very comprehensive review of the performance of any given employee; however, this system is not without its foibles.

As previously stated, one perspective is that supervisors give more accurate ratings as a result of the accountability provided by the face-to-face meeting with the rated employee. However, Klimoski and Inks' (1990) findings seem to contradict the increased accuracy of ratings due to face-to-face appraisal meetings. The authors found that supervisors rated a poorly performing employee higher when anticipating a face-to-face meeting than when feedback was to be given anonymously in writing, or when no feedback was to be given at all. This is alarming in that a method thought to increase accuracy may actually contribute to the distortion of ratings. Regardless of whether face-to-face meetings with affected ratees result in distortion of ratings, multi-source feedback systems have additional raters who would typically not be present during performance appraisal meetings, namely the peers and subordinates. Furthermore, in order to obtain peer and subordinate ratings, these sources are typically granted anonymity. Therefore, they are not accountable to the ratee. It is the introduction of accountability while maintaining an adequate level of anonymity for subordinate raters that will become the focus of this study.

Originally, multisource feedback systems were used almost exclusively for development purposes (London & Smither, 1995). However, there is a growing trend to incorporate multisource systems into administrative decisions (Dalessio, 1998; London, 2001). Bohl (1996) indicated that 22% of companies surveyed used a multisource rating

system, and of these more than 90% of the organizations reported using their system to make administrative decisions. This change in feedback usage has prompted research into how multisource ratings may be affected. According to Greguras, Robie, Schleicher and Goff (2003) subordinate ratings are affected by the purpose of the multisource system. Specifically, the researchers found that more variance was attributed to the target being rated, as opposed to the bias of the rater, when the purpose was for developmental rather than administrative purposes. This is an indication that there is a need to investigate mechanisms to increase the quality and accuracy of ratings when the multisource system is to be used for administrative or a combination of administrative and developmental purposes. Additional research suggests that ratings made for administrative purposes are generally more lenient (Jawahar & Williams, 1997), less variable (Farh, Canella, & Bedeian, 1991), and less accurate (McIntyre, Smith, & Hassett, 1984) than ratings for developmental purposes. It is for these reasons that the current study investigates the accuracy of subordinate ratings in a multisource feedback system used for administrative purposes.

Subordinates rating supervisors is not an ideal situation and requires thought and planning when considering the use of these ratings. It essentially flips the traditional hierarchy of power, with the subordinates affecting the outcomes of their supervisor. There are types of leadership situations where you would not want subordinate ratings to be carried out. Not all supervisors are meant to be liked, for instance a drill sergeant.

The possible role of the consultant in multi-source feedback systems. Accountability is not the same across situations, and the anonymity provided to subordinate raters does not allow for direct accountability to the supervisor. Lerner and

Tetlock (1999) present four constituent parts to accountability as a whole, which assist in challenging the common belief that accountability and anonymity are mutually exclusive. These parts are the mere presence of another, identifiability, evaluation, and reason-giving. These four components are directly applicable to performance appraisal in the traditional sense; however, some of these are differentially applicable to multi-source feedback. First, "mere presence" occurs when the supervisor must meet with the specified employee. Second, the supervisor is clearly identified as the one who provided the ratings. Third, evaluation is the comparison of the ratings provided against the normative ground rules held by the organization. In this fashion there should be a standard for good performance, and an outlined process for appraisal provided by the organization. Fourth, "reason giving" is the explanation for the appraisal, which generally provided by the supervisor to the subordinate. There is reason to believe that by altering the paradigm of multi-source feedback, these guidelines provided by Lerner and Tetlock (1999) can be more applicable, and potentially increase the accuracy of the performance appraisal system.

By keeping subordinates who serve as raters anonymous to the supervisor, some of Lerner and Tetlock's (1999) factors do not apply; namely "mere presence" and identifiability. It would be advantageous to develop a method to keep subordinate raters accountable without identifying them to the supervisor. One method in which this may be accomplished is the introduction of a third party consultant who would act as the reviewer and distributor of the ratings. The subordinate raters would be accountable to the consultant and anonymous to their supervisor. Thus, they would be identifiable to the consultant who could follow-up with the rater if asked to provide reasons for the ratings.

In this way, the four components of accountability posited by Lerner and Tetlock (1999) can be upheld, without the subordinate being identified to anyone who has a stake in the performance appraisal process.

The theoretical significance for the need to have an external consultant oversee multi-source feedback in an effort to improve accuracy comes from the assumption that individuals are motivated to seek and maintain the approval of those to whom they are held accountable (Tetlock, 1985). However, the advantage of a third-party consultant as the reviewer and distributor of the performance ratings goes beyond reducing the amount of inaccuracy (i.e. reducing error due to face-to-face meetings). It may also diminish any long-term effects of low ratings. The consultant is not a figure with whom the ratee will be interacting on a daily basis, and therefore any misgivings may have less of an effect on the daily interaction between an employee and their supervisor.

The two sides to accountability. The differentiation between accountability to a supervisor and accountability to a third party consultant leads to the idea of two separate forms of accountability. The first form is appeasement-accountability, and is characterized by the use of an "acceptability heuristic" (Tetlock, 1985). Following this heuristic, should an employee be accountable to someone with a known view or opinion that the employee wishes to appease, the rating or decision provided will most likely reflect that opinion (Tetlock, 1983; Tetlock, Skitka, Boettger, 1989). Decisions made using the acceptability heuristic generally require very little cognitive depth, relying on information readily available in the environment, and are typically obvious or well known (Tetlock, 1992). For instance, knowing how the supervisor views his or her own performance may influence the subordinate's evaluation to agree with the supervisor's

self-view in order to avoid conflict or confrontation. Thus, the ratings are attenuated towards what the rater thinks the ratee believes to be true, in an effort to avoid confrontation and awkwardness. Ultimately, this type of accountability may make the ratings less accurate.

Alternatively, there is "accuracy-accountability" where being held accountable assists raters to provide ratings that represent the ratees' actual behavior. This may be achieved by keeping the specific views and opinions of the individual to whom the raters are accountable unknown. Whereas with appeasement-accountability the raters engage in shallow cognitive processing, the raters subjected to accuracy-accountability are more likely to engage in an active and attentive search for information, in an attempt to anticipate possible criticism and develop counterarguments for those criticisms (Schlenker, 1986; Tetlock, 1985). Thus, the goal becomes providing ratings that one can most readily defend the accuracy of. In support of this line of thinking, studies show that when accountable to an unknown audience, decision makers show more cognitively complex decision making (McAllister, Mitchell & Beach, 1979; Tetlock & Kim, 1987). In order to accomplish this in the present study, the participants in the accuracy-accountability condition will believe a third party consultant will review their ratings. The third-party consultant has no vested interests and strives to increase the accuracy of the system; hence, their specific view of the ratee's performance will be unknown. These characteristics of the consultant are thought to encourage raters to provide accurate ratings. Thus, there is good reason to believe that by imposing accuracy-accountability through the use of a third-party consultant as the supposed distributor and overseer of

performance ratings, the accuracy of subordinates' ratings of their supervisors could be improved, hence:

Hypothesis 1a: Subordinates who are accountable to the consultant will provide more accurate ratings than those who are not held accountable.

Hypothesis 1b: Subordinates who are accountable to the third party will provide more accurate ratings than those held accountable to the ratee.

Hypothesis 1c: Subordinates who are not held accountable will provide more accurate ratings then those who are accountable to the ratee.

Method

Experimental Sample and Procedure

A total of 108 workers were recruited from Amazon Mechanical Turk (Buhrmester, Kwang, & Gosling, 2011) to serve as participants for the present study. Data were collected online over the course of two sessions during which the participants, respectively, watched then rated videos of four corporate trainers. The participants were led to believe that career services was planning to bring one trainer in as an instructor for a new course offered at the university and that the participants' ratings would be used to help choose this trainer. Furthermore, participants were told that because there were many candidates for this role, they would be rating several trainers.

The first session began with the letter of information (See Appendix A.), informed consent (See Appendix B.), and the manipulation (See Appendix C.). The participants were told that their ratings would not be reviewed by anyone (Condition 1), or that their ratings would be reviewed by the trainers who will be contacting them to justify their ratings (Condition 2), or that their ratings would be reviewed by a third party consultant who would be contacting them to justify their ratings (Condition 3). In order to guarantee participants would believe that they would be contacted they were asked to enter a valid email address. With the purpose of ensuring the security of the participants, the email address was not recorded; only whether or not one was entered.

Participants were told that they would receive three dollars for full completion of the study. They were also told that they would receive this after they had responded to the follow-up questions posed by either the trainers in Condition 2 or the consultant in Condition 3. In the anonymous condition (Condition 1) they were told they would be

compensated after the second session (i.e., after providing their ratings of the candidates). The participants completed a brief survey including their demographics. (See Appendix D.) Following this, the participants viewed four videos of four different trainers. Each video was roughly 7 minutes long. After watching each video the participants were presented with a brief survey asking questions about the video to verify that the participants attended to it. (See Appendix E.) This survey asked specific questions about what happened in the videos.

The second session was the rating session and took place between 24 and 48 hours after the first session. The delay between the observation of the videos and the rating sessions was to simulate memory demands of actual performance appraisals where the ratings are based on memory and not direct observation (e.g. Murphy & Balzer, 1986; Wagner and Goffin, 1997). When logging onto their account the participants were directed to a survey asking them to rate each of the four trainers on a number of skills taking the form of behaviorally anchored rating scales (See Figure 1; Smith & Kendall, 1963) format adapted from McIntyre, Hoover, & Gilbert (1997) (See Appendix F.). In this fashion, participants placed the rating of each trainer on a scale from 0 – 100, based on the displayed behavior. To assist in memory recall a thumbnail image of each trainer was present in the column of each set of scales. Participants were also required to justify their ratings by typing an explanation for the rating provided. Because the anonymous condition (Condition 1) did not have a target for which the rater needed to justify their responses, they completed a questionnaire about judgement preferences instead (See Appendix G.). Following the rating procedure, participants in Conditions 2 and 3

Figure 1.

Example of a Behaviorally Anchored Rating Scale Format.

Trainer A, Performance Dimension 4: Visual Aids.

A high scorer:

- uses visual aids that are appropriate, legible, relevant, and beneficial
- uses visual aids for main points

A low scorer:

- uses too many or too few visual aids
- uses visual aids that are dull
- uses visual aids that are sloppy

VERY HIGH This indicates good use of appropriate visual aids	100	Visual aids are appropriate, legible, relevant, and beneficial
	84	Visual aids are used for main points
	67	
MEDIUM This indicates moderate use of visual aids that are of moderate quality	50	
	33	Too many or too few visual aids are used; Visual aids are dull
VERY LOW This indicates visual aids that are inadequate or poorly utilized	0	Visual aids are sloppy

Please rate the performance of the trainer by using any number from <u>0 to 100</u> (based on the above scale).

Trainer A: _____ (Scores not limited to those presented above)

completed a few more scales to check whether or not they expected to receive an email from the consultant or trainer (See Appendix H.). The participants also completed a series of questions to determine if they adequately remembered the videos from the first session (See Appendix I). Next, the participants viewed the debriefing form, which would indicate that deception was used. In all three conditions, participants were made aware that the university is not actually planning to hire any of the trainers. Furthermore, in Condition 2 the participants were told that they would not be contacted by the trainers nor were their email addresses recorded. Similarly, in Condition 3 the participants were made aware that they would not be contacted by the consultant, nor were their email addresses recorded (See Appendix J.).

Corporate Trainer Videos

The experiment implemented four videos of speakers from a Stanford series of seminars. Each clip was approximately seven minutes in length and was viewed during the first session.

Experimental Manipulations

There were three accountability conditions. Condition 1 was no accountability (control condition). Condition 2 is the appeasement-accountability condition where the participants were held accountable to the trainers, and justified their ratings to the trainers. Condition 3 is the accuracy-accountability condition, where the participants were led to believe they were accountable to a third party consultant who would contact them to justify their ratings.

Dependant Measures

The dependant variable of interest was accuracy. Much of the current research focuses on the four accuracy measures used by Cronbach (1955) which use the true score as a basis for comparison. The true score in this study is the average of scores provided by experts. In this case, the experts were 14 graduate students, nine of which were male and five were female. The expert raters were familiar with the role of trainers and were able to watch the videos any number of times before assigning ratings. Furthermore, to enhance the accuracy of the true scores, the expert raters assigned ratings immediately after viewing the videos, as opposed to the participants who were subject to a 24 to 48 hour delay. This follows standard procedures for developing performance rating true scores (see Jelley & Goffin, 2001; Wagner & Goffin, 1997).

According to Murphy and Cleveland (1995), accuracy has four components which can be measured: elevation accuracy (EL), differential elevation (DE), stereotype accuracy (SA), and differential accuracy (DA) (See Figure 2.). Elevation accuracy is the differential grand mean. Inaccuracy with respect to this component reflects a rater's tendency to rate too high or too low, averaged across all ratees and items, relative to the true score. Differential elevation is the differential main effect of ratees. This reflects the rater's accuracy in differentiating among ratees, averaging across all items and controlling for the rater's level of EL. DE evaluates a rater's accuracy in distinguishing between employees based on their total job performance scores. As such, it is applicable to administrative functions and performance appraisal. Stereotype accuracy (SA) is the differential main effect of the rating scale items averaging across ratees and controlling for the rater's level of EL. This index indicates the accuracy with which a group of ratee's average performance on different items is differentiated by the rater. SA

Figure 2.
Cronbach's (1955) Components of Accuracy

$$EL = \sqrt{(\bar{x}_G - \bar{t}_G)^2}$$

$$DE = \sqrt{\frac{1}{n}\sum_i[(\bar{x}_i - \bar{x}_G) - (\bar{t}_i - \bar{t}_G)]^2}$$

$$SA = \sqrt{\frac{1}{k}\sum_j[(\bar{x}_j - \bar{x}_G) - (\bar{t}_j - \bar{t}_G)]^2}$$

$$DA = \sqrt{\frac{1}{kn}\sum_i\sum_j[(x_{ij} - \bar{x}_i - \bar{x}_j + \bar{x}_G) - (t_{ij} - \bar{t}_i - \bar{t}_j + \bar{t}_G)]^2}$$

n: number of ratees.
k: number of items.
$\bar{x}_G$: rater grand mean (across items and ratees).
$\bar{t}_G$: true score grand mean (across items and ratees).
$\bar{x}_j$: rater mean for the ratee i (across items).
$\bar{t}_j$: true score mean for ratee i (across items).
$\bar{x}_i$: rater mean for item j (across ratees).
$\bar{t}_i$: true score mean for item j (across ratees).
x_{ij}: rating for ratee i on item j.
t_{ij}: true score for ratee i on item j.

describes how accurate the raters are at differentiating between the dimensions. This type of accuracy which would be useful when assessing training needs. Differential accuracy reflects the variance not accounted for by the rater's level of EL or by the main effects of ratees or items. DA is typically interpreted as the accuracy in diagnosing the strengths and weaknesses of individual ratees, thus, relevant for feedback to employees. The present study places significance on DE and DA. This is because the focus of the ratings is administrative in nature and the goal is to differentiate between the trainers.

Statistical Analyses

Preliminary analyses were conducted to be sure that participants attended to the videos and provided purposeful responses in each part of the study. Because the present study was conducted online, a number of safe checks have been put in place to ensure active participation and purposeful responding (See Meade & Craig, 2012). The first check to be completed was those of the post-video questionnaires to ensure the participants did in fact attend to the videos. (See Appendix E.) The second set of checks was to ensure that the participant interpreted the manipulation correctly, which is completed once the participant has assigned ratings, but before they are given the debriefing form (See Appendix H. and I.). To test the hypotheses presented, each of the accuracy components were computed. In order to test for differences between the conditions, a t-test was conducted for each accuracy component. Therefore, 12 t-tests were conducted.

Results

Participants

In total, data from 108 participants were used in the analyses. A check for carless responding did not result in the removal of any participants (Meade & Craig, 2012). The anonymous condition (Condition 1) contained 31 participants, while the condition where participants were accountable to the trainers (Condition 2) and the condition where participants were accountable to the external consultant (Condition 3) contained 37 and 40 respectively. Of the 108 participants 50 were male and 54 were female; the gender information on four participants was missing. The mean age was 34.29 with a standard deviation of 12.37. The minimum age was 19, with a maximum of 68.

In terms of work demographics, all 108 participants had some work experience; 77 participants reported that they were currently employed, with 56 of those in full-time work. Furthermore, 72 participants reported more than five years of work experience in their lifetime, with 54 of those having more than ten years of work experience. A total of 74 participants reported experience in a supervisory role, with 51 reporting between one and five years, and 15 reporting greater than five years. Directly related to performance appraisal, 81 participants reported experience evaluating another employee's performance, and 95 reported having their performance rated.

Preliminary Analyses

Missing data for rating items were replaced using expectation maximization data imputation. This resulted in five value substitutions, which was 0.29% of all items and was limited to three participants. Cronbach's (1955) components of accuracy were computed, based on a 100 point Behaviorally Anchored Rating Scale, using formulae

provided by Sulsky and Balzer (1988), Cardy and Dobbins (1994; see Figure 2). Descriptive statistics for each condition appear in Table 1. The highest possible values for differential elevation (i.e. the most inaccurate) were calculated in order to provide additional meaning to the values seen in Table 1. These mean values were 73.12, 77.55, and 74.84 for Condition 1, Condition 2 , and Condition 3 respectively. The lowest possible values (i.e. the most accurate) would be zero. Accuracy components can be likened to standard deviation values in that the magnitude is based largely on the range of the scale. Accordingly, one would expect larger accuracy component scores from a 100-point BARS scale in comparison to a five-point Likert scale. As mentioned, smaller values on all four components of accuracy are indicative of greater accuracy. As can be seen in Table 1 the means for differential elevation and stereotype accuracy are in the predicted order with Condition 3 being the most accurate and Condition 2 being the least accurate. The pattern changes for elevation and differential accuracy; Condition 1 is the most accurate and Condition 2 is the least accurate.

Test of Hypotheses.

To test the main hypotheses, a series of independent sample t-tests were conducted for each of the four accuracy components (EL, DE, SA, & DA; See Table 2). In support of Hypothesis 1a, the comparison between Condition 1 and Condition 3 was significant for differential elevation $t(69) = 2.27$, $p = .027$. The remaining accuracy components for testing Hypothesis 1a did not reach significance. Partially supporting Hypothesis 1b, the comparison between Condition 2 and Condition 3 was marginally significant for differential elevation $t(75) = 1.73$, $p = .087$. The remaining accuracy

components for testing Hypothesis 1b did not reach significance. Neither DE or DA were significant for Hypothesis 1c, therefore it was not supported.

Table 1.
Cell Means and Standard Deviations for all Accuracy Measures

Dependent Variable	Condition 1 Anonymous		Condition 2 Accountable to Trainer		Condition 3 Accountable to Consultant	
	Mean	SD	Mean	SD	Mean	SD
Elevation (EL)	8.28	5.9	12.6	9.8	10.51	7.84
Differential Elevation (DE)	13.66	5.83	14.23	11.32	10.87	4.54
Stereotype Accuracy (SA)	7.56	2.83	8.96	10.72	7.2	6.91
Differential Accuracy (DA)	9.27	3.76	12.81	20.02	9.55	3.8

Note. SD = Standard Deviation.

Table 2.
Independent Sample t-tests

Comparison	Elevation		Differential Elevation		Stereotype Accuracy		Differential Accuracy	
	T	df	t	df	t	Df	T	Df
Condition 1 & Condition 3	-1.46	69	2.27**	69	0.42	69	-0.31	69
Condition 2 & Condition 3	1.1	75	1.73*	75	0.96	75	1.01	75
Condition 1 & Condition 2	-2.15**	66	-0.25	66	-0.7	66	-0.97	66

Note. 2-tailed, ** p<.05. *p<.10.

Discussion

Multisource feedback systems are being utilized in a growing number of organizational functions. As was previously mentioned the initial purpose of multisource feedback was to assist in the development of employees (London & Smither, 1995). It is believed that ratings from different levels (i.e. supervisor, peer, or subordinate) may provide a unique perspective of a ratee's performance because they have differential exposure to the target's behavior. The result is a comprehensive evaluation of performance, which assists in tasks like goal setting, or the application of training. More recently, however, the function of multisource feedback has shifted to administrative or joint administrative and developmental purposes (Dalessio, 1998; London, 2001). It is this shift that sparked the current research. Information from multisource rating systems is being used to make important decisions; however, the quality of the ratings from subordinates has not been fully investigated.

The concern about accountability and the possible implications it has for accuracy of subordinate ratings was developed from the best practices that are associated with multisource feedback; subordinates are granted anonymity when they provide ratings. Anonymous submission of ratings allows subordinates to rate their supervisors with less fear of possible retaliation if a low rating is warranted. We know that this policy protects subordinates; however, the effect it has on accuracy of ratings has not yet been investigated. This is an example where practice has moved in one direction, and the research to substantiate the practice has not kept pace. The present research is a first attempt to close this aspect of the researcher-practitioner gap.

The goal of the present study was two-fold. The first goal was to apply an accountability structure to subordinate ratings of their supervisors in hopes of increasing the accuracy of subordinate ratings. This task was more complicated than initially thought due to the overwhelming need to ensure rater anonymity. The second goal came out of this need, the differentiation between two types of accountability. As mentioned, the first type is appeasement-accountability, which occurs when a rater must meet face-to-face with, or is identified to, the ratee. Therefore the rater may feel the need to appease the ratee. The premise being that should a subordinate be identified to their supervisor as the rater, the subordinate may want to give a rating that would be accepted by the supervisor in order to avoid possible conflict. This type of accountability would most likely result in an inflation of ratings. Rating inflation would affect rating accuracy if everyone inflated his or her ratings to the same degree and if the rating scale had no upper value limit. However, people do not inflate rating to same degree, and there is a definite ceiling to every rating scale. Consequently, it is believed that along with the inflation of ratings, this type of accountability will decrease accuracy of the ratings. The second type of accountability is accuracy-accountability. As the name implies, this type of accountability should increase accuracy. This is accomplished by holding the rater accountable to a consultant who is only interested in the ratings being as accurate as they can be. In this case, the raters would not be identified to the ratees.

In order to assess the effects of the two types of accountability on performance rating accuracy, the present study compared Cronbach's (1955) accuracy components between a completely anonymous condition (Condition 1), an appeasement-accountability condition (Condition 2), and an accuracy-accountability condition

(Condition 3). The results support the notion that the accuracy-accountability condition provided more accurate ratings than the anonymous and appeasement-accountability conditions for differential elevation. As previously mentioned, differential elevation is interpreted as the accuracy in rank ordering ratees, which is applicable to administrative decisions. The current study was administrative in nature, thus this finding is very promising. However, the comparison between the appeasement-accountability condition and the anonymous condition was not significant for differential elevation.

Limitations

The first limitation in the present study is the use of the Behaviorally Anchored Rating Scales. BARS are viewed as one of the most ideal rating formats currently used in performance management. They provide specific behavioral examples for each level of performance. Furthermore, BARS ratings are considered to be reliable and valid (Stoskopf, Glik, Baker, Ciesla, & Cover, 1992). However, it is possible that because the BARS is very specific about the behaviours associated with levels of performance, there was less "room" for the manipulations to affect the accuracy of the ratings. Furthermore, BARS is very expensive and time consuming to develop. Therefore, it might less used in practice, which affects the generalizability of the present study. As discussed below, future research should consider different scales.

A second limitation would be the use of professional trainers for each of the four videos. Alternatively, if the trainers were selected at random from the population, the raters may be more able to differentiate between those that are high performing versus those that are low performing.

The focus of the present study was on administrative decisions stemming from performance appraisal. Although this is an important area for performance appraisal research, another important goal of multisource feedback systems is employee development. Cleveland, Murphy, and Williams (1989) differentiate between-person decisions and within-person decision when it comes to performance appraisal; between-person decisions being those with administrative purposes, and within-person decisions being those involving development. These types of decisions are thought to be at ends with each other, however many organizations use ratings for both reasons. Furthermore, the accuracy components most relevant to each type of decision are different. The accuracy component most relevant to administrative functions is DE, whereas the component most relevant to development is DA. DE represents a global indication of performance, whereas DA represents the individual strengths and weaknesses of an employee. A study by Goffin & Jelley (2001) found that depending on how the raters were primed the accuracy component that was most affected was different. Thus, raters may alter their rating method based on the purpose of the appraisal. The current study was administrative in nature. This may explain why DE was significant and DA was not. The focus on administrative decisions presents only a portion of accountability of subordinate ratings. Additional research should incorporate the development aspect of performance appraisal.

Another possible limitation is the external consultant and the values which he or she is thought to possess. In order to capitalize on accuracy-accountability a very specific consultant is necessary. This consultant should be committed to having accurate performance ratings. Furthermore, the consultant should know about the industry and

organization, but not be a figure that the employees interact with regularly. This exemplifies the necessity of searching and selecting the most appropriate consultant. One solution would be to use a consulting company that has accuracy as a part of its values or "brand".

Generalizability from the present laboratory experiment to actual performance appraisals in organizations is another possible limitation. We used participants who took part through Amazon Mechanical Turk (Buhrmester, Kwang, & Gosling, 2011). This method also allowed us to access workers from a range of work backgrounds and histories. As previously indicated, 81 of the 108 participants had experience rating another employee's performance, while 95 had their performance rated. Furthermore, the age and work tenure of the participants was diverse. This allows for greater generalizability to the work force than a sample of first-ear undergraduate students. While this is a step in the right direction, there are still unknowns about the participants. For instance, the type of performance appraisal or rating scale that those who have received or given performance ratings used is unknown.

Furthermore, part of the distinction between the appeasement-accountability and accuracy-accountability is the potential for social consequence following the performance appraisal. Appeasement-accountability is thought to inflate ratings in order to avoid retaliation from the ratee. Accuracy-accountability does not encounter this problem because the rater is unknown to the ratee. The use of an online study where the participants have never met the trainers, nor will they, is a limitation. The participants were led to believe that the trainers or consultant would be contacting them; however, the

participants may not have felt the added pressure of needing to work with the trainers on a daily basis.

The nature of accuracy research requiring true score estimation is inherently a limitation to performance appraisal research. In order to obtain true score values, the expert raters must watch the same stimuli as the participants, which limits this type of research to relatively short video clips of job performance. In the present study four such clips were observed, each around seven minutes in length. This is different from applied performance appraisals where the raters observe the ratees's performance in various situations over time. Furthermore, in applied situations, each rater is differentially exposed to the ratee's job performance. However, Bernardin and Villanova (1986) stated that the key to generalizability is psychological fidelity between research and criterion settings, fnot literal similarity. For instance, one method to enhance generalizability is raters should be knowledgeable about the job for which they are rating. Trainers as a target of the performance appraisal were used because most, if not every job requires training from a superior; thus, allowing us to capture a diverse subject pool while making sure that the participants were familiar with the job associated with the ratees.

Furthermore, generalizable aspects of multi-source feedback were included to support the psychological fidelity of the work. Balzer et al. (2004) state that if subordinate ratings are to be included in a performance appraisal system, three to five ratings should be collected. Participants were made aware that their scores would be used in tandem with scores from other participants. Additionally, participants were instructed to submit their ratings between 24 and 48 hours after viewing the videos of the

trainers. This was to ensure that the ratings provided were based on the memory of the raters, as would be the case with actual performance ratings.

Future Research

Future research would benefit from a follow-up study comparing the results from the BARS used in the present study with a more commonly used ratings scale. One of the more common rating scale formats is the Graphic Rating Scale (GRS) (Tziner & Kopelman, 1988). According to Yun, Donahue, Dudley, and McFarland (2005), the GRS format consists of bipolar adjective scales. Furthermore, these scales are somewhat generic in that they tend to be applicable to a number of jobs, which makes them easier to create (Tziner & Kopelman, 2002). However, GRS lack behavioral specificity with the different levels of performance; hence, they can be more difficult to use. Because these scales are less specific in how they describe the behavior, there is potential for raters to interpret the values on the scales differently. There is reason to believe that with more room for interpretation, the GRS may result in greater differences in accuracy between conditions.

Greguras et al. (2003) noted that subordinate ratings are affected by the purpose of a multisource feedback system. A possible follow-up study would apply the appeasement-accountability and accuracy-accountability conditions to a developmental context. It would be predicted that the developmental context may result in fewer differences between conditions because the consequences of poor ratings are not as salient as those in an administrative context. Another line of research would suggest that in a developmental context the goal is to identify strengths and weaknesses of the ratee. This is in line with the accuracy component of differential accuracy. It may be the case

that the accountability conditions differentially affect a rater's ability to identify the ratee's strengths and weaknesses.

The use of analysis of covariance is a possibility for future research. The present study collected data regarding the work experience of the raters. The effects found in the present study may be altered when controlling for some of these variables, such as tenure, supervisory experience, and experience with performance appraisal.

Another area for future research would be to use a field sample to support the findings of the present experimental study. In field settings there are no expert raters to provide true scores, thus the methodology used in the current study and the accuracy components cannot be used. However, O'Neill, Goffin and Gellatly (2012) use a method of variance partitioning through multi-level modelling which can be applied to field data. This method examines the proportion of variance attributable to the ratee, the rater, the interaction between the ratee and rater, and finally random error. O'Neill et al. (2012) describe ratee main effects variance as variance attributable to ratee performance differences and is considered an analogue for true score variance. Rater main effects variance comprises variance involving raters' systematic deviations from the typical ratee rating. Examples of this include leniency or severity. The ratee by rater interaction effects variance involves systematic rating variance associated with dyad-level rater-ratee pairings. In essence, the goal is to maximize ratee main effects variance, and minimize the remaining three sources.

A comparison can be made between two or more groups across all of the above sources of variance. For example, cases can be sorted into high or low familiarity with the ratee. A comparison of the proportion of variance attributable to the ratee should be

higher for the high familiarity group as opposed to the low familiarity group (O'Neill et al, 2012). Similarly, the groups can be divided into experimental manipulations as opposed to familiarity. This provides another means in which the hypotheses in the present study can be tested. The proportion of variance attributable to the ratee should be greater for the accuracy-accountability condition as compared to appeasement-accountability, and the anonymous condition. A field sample would bring the social factors involved in performance appraisal, which may make the manipulations more salient.

Finally, future research should extend the current findings to peer raters. Subordinate raters and peer raters are two additional sources of ratings in a multi-source feedback system. These groups of raters are similar in that they are both typically granted anonymity when providing ratings of an employee. Furthermore, both peer and subordinate raters have the potential for negative consequences if they are identified to the ratee. Along with the possibility for retaliation from ratees that subordinates face, peer raters may also feel a sense of competition with the ratee. In the case of peer raters, both rater and ratee may be on the same level and therefore comparable when considered for rewards like promotion. Consequently, peer raters may feel additional pressures when supplying ratings for a ratee, not all of which may inflate ratings. This would be another area where accountability may assist in ensuring raters are providing accurate ratings.

Conclusions

Based on the results of the present research, it appears that accountability may have an effect on the accuracy of subordinates' ratings of their supervisors for

administrative purposes. It appears that accuracy-accountability provides significantly more accurate ratings than anonymous subordinate ratings according to Cronbach's (1955) component of differential elevation. Similarly, the difference between accuracy-accountability and appeasement-accountability was marginally significant for differential elevation. These findings are very encouraging because they are in the predicted direction. Additionally, differential elevation is especially relevant to administrative functions, which is also in line with the current study. As previously mentioned, this is the first study investigating the accuracy of subordinate ratings under different accountability conditions. Further research should continue to investigate methods to increase accuracy of subordinate ratings.

References

Balzer, W. K., Greguras, G. J., & Raymark, P. H. (2004). In Thomas J. C. (Ed.), *Multisource feedback*. Hoboken, NJ, US: John Wiley & Sons Inc, Hoboken, NJ.

Bernardin, H. J., & Villanova P. (1986). Performance appraisal. In E. Locke (Ed.), Generalizing from laboratory to field settings. Boston: Health/Lexington.

Bohl, D.L. (1996) Minisurvey: 360-degree appraisals yield superior results, survey shows. *Compensation and Benefits Review, 28(5),* 16-20.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*(1), 3-5. doi:http://dx.doi.org/10.1177/1745691610393980

Cardy, R. L., & Dobbins, G. H. (1994). *Performance appraisal: Alternative perspectives*. Cincinnati, Ohio: South-Western Publishing Co.

Cleveland, J. N., Murphy, K. R., & Williams, R. E. (1989). Multiple uses of performance appraisal: Prevalence and correlates.*Journal of Applied Psychology, 74*(1), 130-135. doi:http://dx.doi.org/10.1037/0021-9010.74.1.130

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

Dalessio, A.T. (1998). Using multisource feedback for employee development and personnel decisions. In Smither J.W. (Ed.), *Performance appraisal: State of the art in practice.* San Francisco: Jossey-Bass.

Farh, J., Cannella, A. A., & Bedeian, A. G. (1991). Peer ratings: The impact of purpose on rating quality and user acceptance. *Group & Organization Studies, 16*(4), 367-386.

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

Jawahar, I. M., & Williams, C. R. (1997). Where all the children are above average: The performance appraisal purpose effect. *Personnel Psychology, 50*(4), 905-925.

Klimoski, R., & Inks, L. (1990). Accountability forces in performance appraisal. *Organizational Behavior and Human Decision Processes, 45*(2), 194-208. doi:10.1016/0749-5978(90)90011-W

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin, 125*(2), 255-275. doi:10.1037/0033-2909.125.2.255

London, M. (2001). The great debate: Should multisource feedback be used for administration or development only? In Bracken, D.W., Timmreck, C.W., Church, A.H. (Eds.), *The handbook of multisource feedback: The comprehensive resource for designing and implementing MSF processes, (*pp. 368-385). San Francisco: Jossey-Bass.

London, M., & Smither, J. W. (1995). Can multi-source feedback change perceptions of goal accomplishment, self-evaluations, and performance-related outcomes? theory-

based applications and directions for research. *Personnel Psychology, 48*(4), 803-839.

McAllister, D. W., Mitchell, T. R., & Beach, L. R. (1979). The contingency model for the selection of decision strategies: An empirical test of the effects of significance, accountability, and reversibility.*Organizational Behavior & Human Performance, 24*(2), 228-244. doi:10.1016/0030-5073(79)90027-8

McIntyre, F.S., Hoover, G.A., & Gilbert, F.W. (1997). Educating oral presentations using behaviorally anchored rating scales. *Academy of Educational Leadership Journal*, 1(2).

McIntyre, R. M., Smith, D. E., & Hassett, C. E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. *Journal of Applied Psychology, 69*(1), 147-156.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437-455. doi:http://dx.doi.org/10.1037/a0028085

Mero, N. P., Guidice, R. M., & Brownlee, A. L. (2007). Accountability in a performance appraisal context: The effect of audience and form of accounting on rater response and behavior. *Journal of Management, 33*(2), 223-252. doi:10.1177/0149206306297633

Murphy, K.R., & Cleveland, J.N. (1995). *Understanding performance appraisal.* Thousand Oaks, CA: Sage. Chapter 10 (pp. 267-298)

Murphy, K. R., & Balzer, W. K. (1986). Systematic distortions in memory-based behavior ratings and performance evaluations: Consequences for rating accuracy. *Journal of Applied Psychology, 71*(1), 39-44

O'Neill, T. A., Goffin, R. D., & Gellatly, I. R. (2012). The use of random coefficient modeling for understanding and predicting job performance ratings: An application with field data. *Organizational Research Methods*, *15*(3), 436-462. doi:http://dx.doi.org/10.1177/1094428112438699

Schlenker, B. R. (1986). *Personal accountability: Challenges and impediments in the quest for accountability.* Contract No. 400-77-0099. San Diego, CA: Navy Personnel Research and Development Center.

Schlenker, B.R., Britt, T.W., Pennington, J., Murphy, R., Doherty, K. (1994). The triangle model of responsibility. *Psychological Review,* 101, 632-652.

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*(2), 149-155. doi: 10.1037/h0047060

Stoskopf, C. H., Glik, D. C., Baker, S. L., Ciesla, J. R., & Cover, C. M. (1992). The reliability and construct validity of a behaviorally anchored rating scale used to measure nursing assistant performance. *Evaluation Review*, *16*(3), 333-345.

Retrieved from https://www.lib.uwo.ca/cgi-bin/ezpauthn.cgi/docview/618161372?accountid=15115

Sulsky, L. M., & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology, 73*(3), 497-506. doi:http://dx.doi.org/10.1037/0021-9010.73.3.497

Tetlock, P. E. (1983). Accountability and the perseverance of first impressions. *Social Psychology Quarterly*, *46*(4), 285-292. doi:10.2307/3033716

Tetlock, P. E. (1985). Accountability: The neglected social context of judgment and choice. *Research in Organizational Behavior, 7*, 297-332

Tetlock, P. E. (1992). The impact of accountability on judgment and choice: Toward a social contingency model. In M. P. Zanna (Ed.), (pp. 331-376). San Diego, CA, US: Academic Press, San Diego, CA. doi:10.1016/S0065-2601(08)60287-7

Tetlock, P. E., Kim, J. I. (1987) Accountability and judgment processes in a personality prediction task. *Journal of Personality and Social Psychology*. 52: 700-709.

Tetlock, P. E., Skitka, L., & Boettger, R. (1989). Social and cognitive strategies for coping with accountability: Conformity, complexity, and bolstering. *Journal of Personality and Social Psychology*, *57*(4), 632-640. doi:10.1037/0022-3514.57.4.632

Tziner, A., & Kopelman, R. (1988). Effects of rating format on goal-setting dimensions: A field experiment. *Journal of Applied Psychology, 73*(2), 323-326. doi:http://dx.doi.org/10.1037/0021-9010.73.2.323

Tziner, A., & Kopelman, R. E. (2002). Is there a preferred performance rating format?: A non-psychometric perspective. *Applied Psychology: An International Review, 51*(3), 479-503. doi:http://dx.doi.org/10.1111/1464-0597.00104

Wagner, S. H., & Goffin, R. D. (1997). Differences in accuracy of absolute and comparative performance appraisal methods. *Organizational Behavior and Human Decision Processes, 70*(2), 95-103. doi:10.1006/obhd.1997.2698

Yun, G. J., Donahue, L. M., Dudley, N. M., & McFarland, L. A. (2005). Rater personality, rating format, and social context: Implications for performance appraisal ratings. *International Journal of Selection and Assessment, 13*(2), 97-107. doi:http://dx.doi.org/10.1111/j.0965-075X.2005.00304.x

Appendix A: Letter of Information

Anonymous Condition (Group 1)

Watching and Examining Professional Training Videos: Letter of Information

Project Title: Watching and Examining Professional Training Videos

Investigators: Richard Goffin, Kevin Doyle

Career Services at Western University in Ontario, Canada is offering a seminar course to undergraduate students interested in pursuing a career in business beginning in Winter 2014. The goal of the seminar course is to provide students with perspective on what the transition into the business world will be like. Due to the high number of potential trainers interested in teaching the course the university has asked us to conduct a performance evaluation of the pool of potential trainers using workers who may have received workplace training from their supervisors like you. Your ratings will be used to help determine which trainer will be hired for the upcoming course. You will be providing ratings for four trainers. We would ask that you take on the role of the employee receiving training from a supervisor. Therefore, each trainer can be viewed as a supervisor, whom you are rating. You will receive three dollars for completing the study. In order for your data to be useable the second session must be completed 24 to 48 hours after the first session. If the study is only partially completed, or the second session is not completed between 24 and 48 hours after the first session, you will receive 50 cents instead of the three dollars.

Participation in this study involves two sessions:

a) The first session will take approximately 45 minutes. This session will involve watching 4 videos of trainers and answering a few questions about each.
b) The second session will take approximately 30 minutes and will take place between 24 and 48 hours after the first session has been completed. This session involves providing ratings for the trainers viewed in the first session and a questionnaire upon completing the ratings.

Note: If Session 2 is not completed within 24 to 48 hours after Session 1 your data will not be useable. Therefore, in order to receive the three dollars you must complete Session 2 within 24 and 48 hours of finishing Session 1, otherwise you will receive 50 cents. All surveys are time-stamped to ensure this.

There are no known or discernible risks for your participation in this study.

By signing this consent form, you agree to and understand the following conditions:

1) Participation in this study is voluntary.

2) You may refuse to take part in this study.
3) You may leave the study at any time without loss of promised money associated with the task. Please email (*) to do so. (Note: you will be given money based on which task you are completing. If you do not finish the first session you will not receive money for the second session)
4) You may refuse to answer any question or do any procedure.
5) All information obtained from you will remain confidential.
6) You will receive three dollars for completing the study. Partial completion will result in 50 cents compensation rather than the three dollars.
7) You will be debriefed upon study competition

If you have any further questions about this study, you may contact Kevin Doyle at (*) or Richard Goffin (principal investigator) at (*)

If you have any questions about your rights as a research participant, please contact, Director, Office of Research Ethics at (*).

Appeasement Accountability Condition (Group 2)

Watching and Examining Professional Training Videos: Letter of Information

Project Title: Watching and Examining Professional Training Videos

Investigators: Richard Goffin, Kevin Doyle

Career Services at Western University in Ontario, Canada is offering a seminar course to undergraduate students interested in pursuing a career in business beginning in Winter 2014. The goal of the seminar course is to provide students with perspective on what the transition into the business world will be like. Due to the high number of potential trainers interested in teaching the course the university has asked us to conduct a performance evaluation of the pool of potential trainers using workers who may have received workplace training from their supervisors like you. Your ratings will be used to help determine which trainer will be hired for the upcoming course. You will be providing ratings for four trainers. We would ask that you take on the role of the employee receiving training from a supervisor. Therefore, each trainer can be viewed as a supervisor, whom you are rating. You will receive three dollars for completing the study. In order for your data to be useable the second session must be completed 24 to 48 hours after the first session. If the study is only partially completed, or the second session is not completed between 24 and 48 hours after the first session, you will receive 50 cents instead of the three dollars.

Furthermore, your ratings will be seen by the trainers and they will be contacting you about the ratings you provided. You will be asked to enter your email in order for the trainers to contact you.

Participation in this study involves two sessions:

a) The first session will take approximately 45 minutes. This session will involve watching 4 videos of trainers and answering a few questions about each.
b) The second session will take approximately 30 minutes and will take place between 24 and 48 hours after the first session has been completed. This session involves providing ratings for the trainers viewed in the first session and a questionnaire upon completing the ratings.

Note: If Session 2 is not completed within 24 to 48 hours after Session 1 your data will not be useable. Therefore, in order to receive the three dollars you must complete Session 2 within 24 and 48 hours of finishing Session 1, otherwise you will receive 50 cents. All surveys are time-stamped to ensure this.

There are no known or discernible risks for your participation in this study.

By signing this consent form, you agree to and understand the following conditions:

1) Participation in this study is voluntary.
2) You may refuse to take part in this study.
3) You may leave the study at any time without loss of promised money associated with the task. Please email (*)to do so. (Note: you will be given money based on which task you are completing. If you do not finish the first session you will not receive money for the second session)
4) You may refuse to answer any question or do any procedure.
5) All information obtained from you will remain confidential.
6) You will receive three dollars for completing the study. Partial completion will result in 50 cents compensation rather than the three dollars.
7) You will be debriefed upon study completion.

If you have any further questions about this study, you may contact Kevin Doyle at (*) or Richard Goffin (principal investigator) at (*)

If you have any questions about your rights as a research participant, please contact, Director, Office of Research Ethics at (*).

Accuracy Accountability Condition (Group 3)

Watching and Examining Professional Training Videos: Letter of Information

Project Title: Watching and Examining Professional Training Videos

Investigators: Richard Goffin, Kevin Doyle

Career Services at Western University in Ontario, Canada is offering a seminar course to undergraduate students interested in pursuing a career in business beginning in Winter 2014. The goal of the seminar course is to provide students with perspective on what the transition into the business world will be like. Due to the high number of potential trainers interested in teaching the course the university has asked us to conduct a performance evaluation of the pool of potential trainers using workers who may have received workplace training from their supervisors like you. Your ratings will be used to help determine which trainer will be hired for the upcoming course. You will be providing ratings for four trainers. We would ask that you take on the role of the employee receiving training from a supervisor. Therefore, each trainer can be viewed as a supervisor, whom you are rating. You will receive three dollars for completing the study. In order for your data to be useable the second session must be completed 24 to 48 hours after the first session. If the study is only partially completed, or the second session is not completed between 24 and 48 hours after the first session, you will receive 50 cents instead of the three dollars.

Furthermore, your ratings will be seen by a consultant who is overseeing the project. The consultant will contact you about the ratings provided. You will be asked to enter your email address in order for consultant to contact you.

Participation in this study involves two sessions:

a) The first session will take approximately 45 minutes. This session will involve watching 4 videos of trainers and answering a few questions about each.
b) The second session will take approximately 30 minutes and will take place between 24 and 48 hours after the first session has been completed. This session involves providing ratings for the trainers viewed in the first session and a questionnaire upon completing the ratings.

Note: If Session 2 is not completed within 24 to 48 hours after Session 1 your data will not be useable. Therefore, in order to receive the three dollars you must complete Session 2 within 24 and 48 hours of finishing Session 1, otherwise you will receive 50 cents. All surveys are time-stamped to ensure this.

There are no known or discernible risks for your participation in this study.

By signing this consent form, you agree to and understand the following conditions:

1) Participation in this study is voluntary.
2) You may refuse to take part in this study.
3) You may leave the study at any time without loss of promised money associated with the task. Please email (*)to do so. (Note: you will be given money based on which task you are completing. If you do not finish the first session you will not receive money for the second session)
4) You may refuse to answer any question or do any procedure.
5) All information obtained from you will remain confidential.
6) You will receive three dollars for completing the study. Partial completion will result in 50 cents compensation rather than the three dollars.
7) You will be debriefed upon study completion.

If you have any further questions about this study, you may contact Kevin Doyle at (*) or Richard Goffin (principal investigator) at (*)

If you have any questions about your rights as a research participant, please contact, Director, Office of Research Ethics at (*).

Appendix B: Informed Consent

Watching and Examining Professional Training Videos: Consent Form

I have read the Letter of Information, have had the nature of the study explained to me, and agree to participate. All of my questions have been answered to my satisfaction.

I, _____, agree to take part in this study.

[print name]

_____ _____

Signature Date (mm/dd/yyyy)

Experimenter's signature:_____ Date:_____

Appendix C: Manipulation

Anonymous Condition/Group 1

Career Services at Western University in Ontario, Canada is offering a seminar course to undergraduate students interested in pursuing a career in business beginning in Winter 2014. The goal of the seminar course is to provide students with perspective on what the transition into the business world will be like. Due to the high number of potential trainers interested in teaching the course the university has asked us to conduct a performance evaluation of the pool of potential trainers using workers who may have received workplace training from their supervisors like you. Your ratings will be used to help determine which trainer will be hired for the upcoming course. You will be providing ratings for four trainers. We would ask that you take on the role of the employee receiving training from a supervisor. Therefore, each trainer can be viewed as a supervisor, whom you are rating. You will receive three dollars for completing the study. In order for your data to be useable the second session must be completed 24 to 48 hours after the first session. If the study is only partially completed, or the second session is not completed between 24 and 48 hours after the first session, you will receive 50 cents instead of the three dollars.

Appeasement Accountability Condition/Group 2

Career Services at Western University in Ontario, Canada is offering a seminar course to undergraduate students interested in pursuing a career in business beginning in Winter 2014. The goal of the seminar course is to provide students with perspective on what the transition into the business world will be like. Due to the high number of potential trainers interested in teaching the course the university has asked us to conduct a performance evaluation of the pool of potential trainers using workers who may have received workplace training from their supervisors like you. Your ratings will be used to help determine which trainer will be hired for the upcoming course. You will be providing ratings for four trainers. We would ask that you take on the role of the employee receiving training from a supervisor. Therefore, each trainer can be viewed as a supervisor, whom you are rating. You will receive three dollars for completing the study. In order for your data to be useable the second session must be completed 24 to 48 hours after the first session. If the study is only partially completed, or the second session is not completed between 24 and 48 hours after the first session, you will receive 50 cents instead of the three dollars.

Furthermore, your ratings will be seen by the trainers and they will be contacting you about the ratings you provided. You will be asked to enter your email in order for the trainers to contact you.

Accuracy Accountability Condition/Group 3

Career Services at Western University in Ontario, Canada is offering a seminar course to undergraduate students interested in pursuing a career in business beginning in Winter 2014. The goal of the seminar course is to provide students with perspective on what the transition into the business world will be like. Due to the high number of potential trainers interested in teaching the course the university has asked us to conduct a performance evaluation of the pool of potential trainers using workers who may have received workplace training from their supervisors like you. Your ratings will be used to help determine which trainer will be hired for the upcoming course. You will be providing ratings for four trainers. We would ask that you take on the role of the employee receiving training from a supervisor. Therefore, each trainer can be viewed as a supervisor, whom you are rating. You will receive three dollars for completing the study. In order for your data to be useable the second session must be completed 24 to 48 hours after the first session. If the study is only partially completed, or the second session is not completed between 24 and 48 hours after the first session, you will receive 50 cents instead of the three dollars.

Furthermore, your ratings will be seen by a consultant who is overseeing the project. The consultant will contact you about the ratings provided. You will be asked to enter your email address in order for consultant to contact you.

Appendix: D: Demographic Information

Participant Information Form

Please enter your first and last initial followed by the day of the month you were born on: i.e. John Smith born on April 23rd would be JS23.: _____

Please enter your email in order for the **trainers/consultant** to contact you: _____

Gender:

- ○ Male
- ○ Female

Age: _____

Please enter the State in which you reside: _____

Please enter the City in which you reside: _____

Primary Language: _____

Are you currently employed?

- ○ No
- ○ Yes
  - ➢ If yes, how many hours?
    - ○ employed full-time (25 or more hours per week)
    - ○ employed part-time (10 - 24 hours per week)
    - ○ employed part-time (9 hours or fewer per week)
    - ○ not employed (studying, travelling, etc.)

How long have you been employed in your lifetime?

- ○ Under 1 year
- ○ 1 to 5 years
- ○ 5 to 10 years

- 10 to 15 years
- 15+ years

Please select the industry that best reflects your work experience.

- Accommodation and food services
- Administrative and support services
- Agriculture, forestry, fishing, and hunting
- Arts, entertainment, and recreation
- Construction
- Educational services
- Finance and insurance
- Government
- Healthcare and social assistance
- Information
- Management of companies and enterprises
- Manufacturing
- Mining, quarrying, and oil and gas extraction
- Professional, scientific, and technical services
- Real estate and rental and leasing
- Retail trade
- Self-employed
- Transportation and warehousing
- Utilities
- Other
  - If other, please specify? _____

Have you ever worked in a supervisory role?

- No
- Yes
  - If yes, for how long?

    - Under 1 year
    - 1 to 5 years
    - 5 – 10 years
    - 10+ years

Have you ever had to evaluate an employee's job performance or give job performance feedback?

- ○ Yes
- ○ No

Have you ever had your performance rated at a job that you've held?

- ○ Yes
- ○ No

Appendix E: Session 1 Video Content Questions

Trainer A

Question #1	Some audience members raised their hands in response to one of the trainer's inquiries.
	True
	False

Question #2	To illustrate a point, the trainer drank a can of Coca Cola.
	True
	False

Question #3	The story has many older versions. The American version from the 1800s involves:
	a) A steam train
	b) **A carriage**
	c) A dog
	d) A bicycle

Question #4	What are the 7 hints to make your ideas stick as successfully as urban legends?
	a) Complex, Specific, Narrative, Emotionless, Self-Explanatory, Comprehensive, Fun
	b) **Simple, Unexpected, Concrete, Credentialed, Emotional, Story that Sticks**
	c) Simple, Resonant, Amusing, Extraordinary, Emotional, Inspirational, Words to live by
	d) Adventurous, Fun, Simple, Exciting, Inspirational, Comprehensive, Illustrative

Question #5	To illustrate a point, the trainer showed the audience pictures of his dog.
	True
	False

Trainer B

Question #1	The trainer argued that teams:
	a) Provide structure for speed but not learning
	b) Provide structure for learning but not speed
	c) Provide structure for both speed and learning
	d) Do not provide structure for speed nor learning

Question #2	To illustrate several key points, the trainer used colour slides.
	True
	False

Question #3	According to the trainer, what is one of the major advantages of using teams?
	a) Teams are more efficient than individuals working on their own
	b) It's easier to motivate a team than it is an individual
	c) **Teams create a structure for people with different perspectives to interact**
	d) Employees are more satisfied working in teams than by themselves

Question #4	A member of the audience asked the trainer a question while he was talking.
	True
	False

Question #5	The trainer indicated that teams are best viewed as a vehicle to get work done.
	True
	False

Trainer C

Question #1	According to the trainer, what is the primary driver of technological change in most industries?
	a) Increasing international competition
	b) Increasing local competition
	c) Demand for more sophisticated products
	d) All of the above

Question #2	To illustrate one of the trainer's points, cartoon animations were included on one of the slides.
	True
	False

Question #3	During his presentation, the trainer mentioned that they had studied the cement industry.
	True

	False

Question #4	The trainer indicated that when trying to introduce change, there is inevitable resistance.
	True
	False

Question #5	To add humour, the trainer's last slide reads "What proportion of an MBA class would voluntarily eat a(n) _____?"
	a) **Earthworm**
	b) Jalapeno
	c) Frog
	d) Spider

Trainer D

Question #1	What is the trainer's main occupation?
	a) Executive
	b) **University Professor**
	c) Manager
	d) Human Resources Practitioner

Question #2	The trainer indicated that the principles of persuasion are universal and

	apply beyond the work context.
	True
	False

Question #3	The trainer joked that the most resistant of all audiences is:
	a) Your friends
	b) Your neighbours
	c) Your coworkers
	d) **Your children**

Question #4	According to the trainer, persuasion/influence is really an art, not a science.
	True
	False

Question #5	To illustrate a point, the trainer asked an audience member to come on stage.
	True
	False

Appendix F: Behaviorally Anchored Rating Scales

Performance Ratings

For this section you will be asked to rate the performance of each trainer. To put this in perspective, you can consider teachers and professors to be trainers.

Consider the following hypothetical example for the following trainer:
Please rate trainer E on how effective he is in the *allocation of time*.
We have listed example behaviours that reflect low to high performance in *allocation of time*, and you should refer to these behaviours as a guideline when evaluating the trainers. However, it is not absolutely necessary for the trainer to engage in a specific behaviour in order to receive the level of performance associated with it. The behaviours just provide an indication of the performance that could be expected at that level.

Trainer E, Performance Dimension Example: Allocation of Time.

VERY HIGH Proper usage of time	100	Budgets time wisely to cover all major points
	84	Time is well allocated to all major facets of the topic
MEDIUM Moderately effective usage of time	50	Allocation of time is acceptable
	33	Too much or too little time is spent on one section
VERY LOW Poor usage of time	0	Presentation is *much* longer than allotted time or finishes *much* too quickly

Please rate the performance of the trainer by using any number from 0 to 100 (based on the above scale).
Trainer E: 54

In the hypothetical example, trainer E's score is 54. This indicates that his performance corresponds fairly closely, but is slightly better than "allocation of time is acceptable". His performance was judged to be lower than what you would associate with "time is well allocated to all major facets of the topic."

Trainer A, Performance Dimension 1: Organization/Preparation.

A high scorer:

- is well organized and audience can easily follow major points
- has major points that follow a logical flow and develop a theme; transitions between points are smooth
- presents consistent and nonredundant info
- begins with an introduction, overview, and outline of the presentation

A low scorer:

- presents too many facts
- does not maintain continuity (eg. excessive pauses or obvious breaks occur)
- appears disorganized and unprepared

VERY HIGH This indicates that the presentation is well structured and organized	100	Presentation is well organized and audience can easily follow major points
	84	Major points follow logical flow and develop theme; Transitions between points are smooth
	67	Info is consistent and not redundant; Presentation begins with intro, overview, and outline
MEDIUM This indicates moderate structure and organization		
	33	Too many facts are presented
	16	Presentation lacks continuity (eg. excessive pauses or obvious breaks occur)
VERY LOW This indicates a poorly structured presentation and organization	0	Presentation appears disorganized and unprepared

Please rate the performance of the trainer by using any number from <u>0 to 100</u> (based on the above scale).

Trainer A: _____ (Scores not limited to those presented above)

Trainer A, Performance Dimension 2: Physical delivery.

A high scorer:

- faces and addresses audience directly
- uses eye contact effectively, appears confident
- uses pleasant facial expressions
- uses effective gestures and body movement
- clothing is appropriate
- commands/keeps audience attention
- appropriately directs audience to the visual aids

A low scorer:

- exhibits nervous mannerisms
- uses little or no eye contact, turns back to audience
- exhibits negative attitude

VERY HIGH This indicates a high level of physical delivery skill	100	Faces and addresses audience; Uses eye contact effectively; Appears confident, using pleasant facial expressions
	84	Uses effective gestures and body movement; Clothing is appropriate; Commands/keeps audience attention
	67	Appropriately directs audience to the visual aids
MEDIUM This indicates a moderate level of physical delivery skill	50	Eyes roam around room; Appears tense or anxious; Constantly paces while talking; Blocks visual aids
	33	Exhibits nervous mannerisms
	16	Little or no eye contact, turns back to audience
VERY LOW This indicates a low level of physical delivery skill	0	Exhibits negative attitude

Please rate the performance of the trainer by using any number from <u>0 to 100</u> (based on the above scale).

Trainer A: _____ (Scores not limited to those presented above)

Trainer A, Performance Dimension 3: Vocal Delivery.

A high scorer:

- clearly pronounces words, projects voice, and uses correct grammar
- uses vivid language, exhibits strong vocabulary
- uses a rate of speech that is varied and appropriate
- has a pleasant pitch and level of formality appropriate for audience

A low scorer:

- has poor speech skills; hesitates while talking, uses “uhs”, “ums”, and “ahs”
- uses poor choice of words; uses too much slang
- mumbles or cannot be heard

VERY HIGH A high level of vocal delivery skill	100	Clearly pronounces words, projects voice, and uses correct grammar
	84	Language is vivid, exhibiting strong vocabulary; Rate of speech is varied and appropriate
	67	Pleasant pitch and level of formality appropriate for audience
MEDIUM This indicates a moderate level of delivery skill	50	Sometimes speaks too loudly or too slow/fast; Sometimes varies tone of voice too much
	33	Poor speech skills; Hesitates while talking, uses too many uhs, ums, and ahs
	16	Uses poor choice of words; Uses too much slang
VERY LOW A low level of vocal delivery skill	0	Mumbles or cannot be heard

Please rate the performance of the trainer by using any number from <u>0 to 100</u> (based on the above scale).

Trainer A: _____ (Scores not limited to those presented above)

Trainer A, Performance Dimension 4: Visual Aids.

A high scorer:

- uses visual aids that are appropriate, legible, relevant, and beneficial
- uses visual aids for main points

A low scorer:

- uses too many or too few visual aids
- uses visual aids that are dull
- uses visual aids that are sloppy

VERY HIGH This indicates good use of appropriate visual aids	100	Visual aids are appropriate, legible, relevant, and beneficial
	84	Visual aids are used for main points
	67	
MEDIUM This indicates moderate use of visual aids that are of moderate quality	50	
	33	Too many or too few visual aids are used; Visual aids are dull
VERY LOW This indicates visual aids that are inadequate or poorly utilized	0	Visual aids are sloppy

Please rate the performance of the trainer by using any number from <u>0 to 100</u> (based on the above scale).

Trainer A: _____ (Scores not limited to those presented above)

Appendix G: Judgement Preferences Scale

Judgment Preferences Scale

This scale examines how you make judgments about a variety of things. Each question is on a scale from 1 (which means strongly disagree) to 7 (which means strongly agree). Simply choose the answer that best describes you.

An example of how to use the scale:

1	*2*	*3*	*4*	*5*	*6*	*7*
Strongly Disagree	*Moderately Disagree*	*Somewhat Disagree*	*Neither Agree nor Disagree*	*Somewhat Agree*	*Moderately Agree*	*Strongly Agree*

---

1. After giving a presentation to a class or a group of people, I would think about how my performance compared to other presenters' performances

   1 2 3 4 5 6 7

2. When considering how sociable I am at a party, I compare my social behavior to other people's social behavior.

   1 2 3 4 5 6 7

3. When considering my romantic relationships, I compare them to my earlier romantic relationships and other people's relationships.

   1 2 3 4 5 6 7

4. My friends' work ethic is difficult to evaluate without thinking of other people's work ethic.

   1 2 3 4 5 6 7

5. When thinking about how friendly my friends are, I think of how friendly other people are.

   1 2 3 4 5 6 7

6. When considering how principled my parents are, I think of how principled other parents are.

   1 2 3 4 5 6 7

7. I compare how loved ones are doing compared to how well other people I know are doing.

1 2 3 4 5 6 7

8. I judge the level of creativity of my work against the level of creativity of other people's work.

1 2 3 4 5 6 7

9. When considering how ambitious I am, I think of how ambitious other people are.

1 2 3 4 5 6 7

10. To get a sense of how committed I am to my religious beliefs, I think about how committed other people are to their religious beliefs

1 2 3 4 5 6 7

11. When evaluating dating partners, I compare their characteristics (for example, kindness or ambition) to the characteristics of other people.

1 2 3 4 5 6 7

12. To assess the creative contribution of coworkers or classmates, I consider the creativity of other people I know.

1 2 3 4 5 6 7

13. When I think of how impressive my family members' goals are, I think of how impressive other people's goals are.

1 2 3 4 5 6 7

14. I think of how politically active people are when considering how politically active I am.

1 2 3 4 5 6 7

15. I consider how liberal or conservative my family is compared to how liberal or conservative other families are.

1 2 3 4 5 6 7

16. When meeting a new coworker or classmate, I compare how outgoing or fun they are to how outgoing or fun other people I know are.

1 *2* *3* *4* *5* *6* *7*

17. When considering how ambitious my educational goals are, I think about what other peoples' educational goals are like.

1 *2* *3* *4* *5* *6* *7*

Appendix H: Post-Experiment Questionnaire

Anonymous Condition (Group 1)

Post Experiment Questionnaire

We have a couple of concluding questions now that you have completed the research.

1. On average, how much attention did you pay when you watched the videos? *Please be honest, these ratings will not influence whether you receive compensation.* (circle your response).

1	2	3	4	5
No attention	Not much attention	Some attention	A great deal of attention	Maximum attention

2. Were any of the instructions in this research difficult to follow?

- ○ Yes
- ○ No

➢ If yes, please tell me which and why?

3. The ratings from this session will provide useful feedback to the trainers (circle your response).

1	2	3	4	5
Strongly Disagree	Disagree	Not sure	Agree	Strongly Agree

4. My thoughts about how the trainers would react affected the ratings I provided.

1	2	3	4	5

Strongly Disagree	Disagree	Not sure	Agree	Strongly Agree

5. I felt accountable to the trainers.

1	2	3	4	5
Strongly Disagree	Disagree	Not sure	Agree	Strongly Agree

6. I am confident that the ratings of the trainers' performance that I provided are accurate

1	2	3	4	5
Strongly Disagree	Disagree	Not sure	Agree	Strongly Agree

7. I took notes during the first session to assist with the second session.

- ○ Yes
- ○ No

Appeasement Accountability Condition (Group 2)

Post Experiment Questionnaire

We have a couple of concluding questions now that you have completed the research.

1. On average, how much attention did you pay when you watched the videos? *Please be honest, these ratings will not influence whether you receive compensation* (circle your response).

1	2	3	4	5
No attention	Not much attention	Some attention	A great deal of attention	Maximum attention

2. Were any of the instructions in this research difficult to follow?

- ○ Yes
- ○ No

- ➢ If yes, please tell me which and why?

_____

_____

_____

_____

_____

3. The ratings from this session will provide useful feedback to the trainers (circle your response).

1	2	3	4	5
Strongly Disagree	Disagree	Not sure	Agree	Strongly Agree

4. My thoughts about how the trainers would react affected the ratings I provided.

1	2	3	4	5

Strongly Disagree	Disagree	Not sure	Agree	Strongly Agree

5. I felt accountable to the trainers.

1	2	3	4	5
Strongly Disagree	Disagree	Not sure	Agree	Strongly Agree

6. I am confident that the ratings of the trainers' performance that I provided are accurate

1	2	3	4	5
Strongly Disagree	Disagree	Not sure	Agree	Strongly Agree

7. I took notes during the first session to assist with the second session.

- ○ Yes
- ○ No

Accuracy Accountability Condition (Group 3)

Post Experiment Questionnaire

We have a couple of concluding questions now that you have completed the research.

1. On average, how much attention did you pay when you watched the videos? *Please be honest, these ratings will not influence whether you receive compensation* (circle your response).

1	2	3	4	5
No attention	Not much attention	Some attention	A great deal of attention	Maximum attention

2. Were any of the instructions in this research difficult to follow?

- ○ Yes
- ○ No

- ➢ If yes, please tell me which and why?

_____

_____

_____

_____

_____

3. The ratings from this session will provide useful feedback to the trainers (circle your response).

1	2	3	4	5
Strongly Disagree	Disagree	Not sure	Agree	Strongly Agree

4. My thoughts about how the consultant reviewing the ratings would react affected the ratings I provided.

1	2	3	4	5
Strongly Disagree	Disagree	Not sure	Agree	Strongly Agree

5. I felt accountable to the trainers.

1	2	3	4	5
Strongly Disagree	Disagree	Not sure	Agree	Strongly Agree

6. I felt accountable to the consultant reviewing the ratings.

1	2	3	4	5
Strongly Disagree	Disagree	Not sure	Agree	Strongly Agree

7. I am confident that the ratings of the trainers' performance that I provided are accurate

1	2	3	4	5
Strongly Disagree	Disagree	Not sure	Agree	Strongly Agree

8. I took notes during the first session to assist with the second session.

- ○ Yes
- ○ No

Appendix I: Session 2 Video Content Questions

Trainer A

Question #1	What is the correct name of the story that the trainer tells in order to illustrate his message?
	a) The Funhouse Dummy
	b) Technology of Today
	c) The Secret to Success
	d) **The Vanishing Hitchhiker**

Trainer B

Question #1	What main topic does the trainer speak about?
	a) **Teams in organizations**
	b) How to keep workers happy
	c) Military psychology
	d) Green initiative in the workplace

Trainer C

Question #1	What is the main topic of the trainer's presentation?
	a) **Innovation and Change**
	b) Success
	c) Economics
	d) Money

Trainer D

Question #2	What is the focus of this trainer's presentation?

	a) Selecting high performing employees
	b) Corporate Responsibility
	c) Capitalism
	d) Persuasion and Influence

Appendix J: Debriefing From

Anonymous Condition (Group 1)

Watching and Examining Professional Training Videos: Debriefing

What were we actually studying? In some work contexts performance appraisal may include a supervisory rating, a self-rating, as well as a number of peer and/or subordinate ratings. It has been suggested that ratings from different levels may provide different perspectives on the performance of any given employee, which can help guide the development and improvement process. When performance ratings are collected from an employee's subordinates, those providing the ratings are granted anonymity from the supervisor of the ratings. The purpose of the present study was to investigate the effect of differences between this completely anonymous condition, accountability to the supervisor, and accountability to an external third party on the accuracy of the ratings provided by subordinates.

Originally, we explained that the purpose of this study was to evaluate the performance of trainers on a series of videos to determine whether Western University would hire them for an undergraduate seminar course. In fact, we are not working with Career Services nor providing your feedback to the trainers.

Why did we use deception? We regret its use but we needed to ensure that you believed the results of these ratings carried some real weight in terms of affecting the outcome of an individual. In the real world, employees' promotions, pay, and even employment status are on the line when they are evaluated. Consequently, employers face significant pressure to ensure their ratings are both accurate and fair, and *should* take their evaluations seriously. If you knew your ratings had no consequences for the trainers, they would be less generalizable to actual performance ratings in real organizations.

We ask that you do not discuss the nature of this study with people who have not already participated. If future participants already know what we are examining, it could potentially have negative effects on our results.

For more information, you may wish to read:

Frink, D. D., & Klimoski, R. J. (2004). Advancing accountability theory and practice: Introduction to the human resource management review special edition. *Human Resource Management Review, 14*(1), 1-17. doi:10.1016/j.hrmr.2004.02.001

Mero, N. P., Guidice, R. M., & Brownlee, A. L. (2007). Accountability in a performance appraisal context: The effect of audience and form of accounting on rater response and behavior. *Journal of Management, 33*(2), 223-252. doi:10.1177/0149206306297633

If you have any further questions about this study, you may contact Kevin Doyle by email at (*).

Appeasement Accountability Condition (Group 2)

Watching and Examining Professional Training Videos: Debriefing

What were we actually studying? In some work contexts performance appraisal may include a supervisory rating, a self-rating, as well as a number of peer and/or subordinate ratings. It has been suggested that ratings from different levels may provide different perspectives on the performance of any given employee, which can help guide the development and improvement process. When performance ratings are collected from an employee's subordinates, those providing the ratings are granted anonymity from the supervisor of the ratings. The purpose of the present study was to investigate the effect of differences between this completely anonymous condition, accountability to the supervisor, and accountability to an external third party on the accuracy of the ratings provided by subordinates.

Originally, we explained that the purpose of this study was to evaluate the performance of trainers on a series of videos to determine whether Western University would hire them for an undergraduate student seminar course. Furthermore, we told you the trainers would be contacting you to discuss the ratings you provided. In fact, we are not working with Career Services, nor providing your feedback to the trainers, nor will the trainers be contacting you. We did not actually record your email at the beginning of Session 1. We simply verified that one was entered.

Why did we use deception? We regret its use but we needed to ensure that you believed the results of these ratings carried some real weight in terms of affecting the outcome of an individual. In the real world, employees' promotions, pay, and even employment status are on the line when they are evaluated. Consequently, employers face significant pressure to ensure their ratings are both accurate and fair, and *should* take their evaluations seriously. If you knew your ratings had no consequences for the trainers, they would be less generalizable to actual performance ratings in real organizations.

We ask that you do not discuss the nature of this study with people who have not already participated. If future participants already know what we are examining, it could potentially have negative effects on our results.

For more information, you may wish to read:

Frink, D. D., & Klimoski, R. J. (2004). Advancing accountability theory and practice: Introduction to the human resource management review special edition. *Human Resource Management Review, 14*(1), 1-17. doi:10.1016/j.hrmr.2004.02.001

Mero, N. P., Guidice, R. M., & Brownlee, A. L. (2007). Accountability in a performance appraisal context: The effect of audience and form of accounting on rater response and behavior. *Journal of Management*, *33*(2), 223-252. doi:10.1177/0149206306297633

If you have any further questions about this study, you may contact Kevin Doyle by email at (*).

Accuracy Accountability Condition (Group 3)

Watching and Examining Professional Training Videos: Debriefing

What were we actually studying? In some work contexts performance appraisal may include a supervisory rating, a self-rating, as well as a number of peer and/or subordinate ratings. It has been suggested that ratings from different levels may provide different perspectives on the performance of any given employee, which can help guide the development and improvement process. When performance ratings are collected from an employee's subordinates, those providing the ratings are granted anonymity from the supervisor of the ratings. The purpose of the present study was to investigate differences between this completely anonymous condition, accountability to the supervisor, and accountability to an external third party on the accuracy of the ratings provided by subordinates.

Originally, we explained that the purpose of this study was to evaluate the performance of trainers on a series of videos to determine whether Western University would hire them for an undergraduate seminar course. Furthermore, we told you the consultant running the investigation would be contacting you to discuss the ratings you provided. In fact, we are not working with Career Services, nor providing your feedback to the trainers, nor will the consultant be contacting you. We did not actually record your email at the beginning of Session 1. We simply verified that one was entered.

Why did we use deception? We regret its use but we needed to ensure that you believed the results of these ratings carried some real weight in terms of affecting the outcome of an individual. In the real world, employees' promotions, pay, and even employment status are on the line when they are evaluated. Consequently, employers face significant pressure to ensure their ratings are both accurate and fair, and *should* take their evaluations seriously. If you knew your ratings had no consequences for the trainers, they would be less generalizable to actual performance ratings in real organizations.

We ask that you do not discuss the nature of this study with people who have not already participated. If future participants already know what we are examining, it could potentially have negative effects on our results.

For more information, you may wish to read:

Frink, D. D., & Klimoski, R. J. (2004). Advancing accountability theory and practice: Introduction to the human resource management review special edition. *Human Resource Management Review, 14*(1), 1-17. doi:10.1016/j.hrmr.2004.02.001

Mero, N. P., Guidice, R. M., & Brownlee, A. L. (2007). Accountability in a performance appraisal context: The effect of audience and form of accounting on rater response and behavior. *Journal of Management*, *33*(2), 223-252. doi:10.1177/0149206306297633

If you have any further questions about this study, you may contact Kevin Doyle by email at (*).

Appendix K: Ethics Approval

0

Western

Department of Psychology

Use of Human Subjects - Ethics Approval Notice

Review Number	**13 01 12**	**Approval Date**	**13 01 29**
Principal Investigator	**Rick Goffin/Kevin Doyle**	**End Date**	**14 01 28**
Protocol Title	**Watching and examining professional training videos**		
Sponsor	**n/a**		

This is to notify you that The University of Western Ontario Department of Psychology Research Ethics Board (PREB) has granted expedited ethics approval to the above named research study on the date noted above.

The PREB is a sub-REB of The University of Western Ontario's Research Ethics Board for Non-Medical Research Involving Human Subjects (NMREB) which is organized and operates according to the Tri-Council Policy Statement and the applicable laws and regulations of Ontario. (See Office of Research Ethics web site: http://www.uwo.ca/research/ethics/)

This approval shall remain valid until end date noted above assuming timely and acceptable responses to the University's periodic requests for surveillance and monitoring information.

During the course of the research, no deviations from, or changes to, the protocol or consent form may be initiated without prior written approval from the PREB except when necessary to eliminate immediate hazards to the subject or when the change(s) involve only logistical or administrative aspects of the study (e.g. change of research assistant, telephone number etc). Subjects must receive a copy of the information/consent documentation.

Investigators must promptly also report to the PREB:
a) changes increasing the risk to the participant(s) and/or affecting significantly the conduct of the study;
b) all adverse and unexpected experiences or events that are both serious and unexpected;
c) new information that may adversely affect the safety of the subjects or the conduct of the study.

If these changes/adverse events require a change to the information/consent documentation, and/or recruitment advertisement, the newly revised information/consent documentation, and/or advertisement, must be submitted to the PREB for approval.

Members of the PREB who are named as investigators in research studies, or declare a conflict of interest, do not participate in discussion related to, nor vote on, such studies when they are presented to the PREB.

Clive Seligman Ph.D.

Chair, Psychology Expedited Research Ethics Board (PREB)

The other members of the 2012-2013 PREB are: Mike Atkinson (Introductory Psychology Coordinator), Rick Goffin, Riley Hinson Albert Katz (Department Chair), Steve Lupker, and TBA (Graduate Student Representative)

CC: UWO Office of Research Ethics

This is an official document. Please retain the original in your files

Curriculum Vitae

Name:	Kevin Doyle
Post-secondary Education and Degrees:	University of Western Ontario London, Ontario, Canada 2005-2010 B.A.
Honours and Awards:	Social Science and Humanities Research Council (SSHRC) Masters Scholarship 2012-2013 Province of Ontario Graduate Scholarship (OGS) Doctoral Scholarship 2013-2014
Related Work Experience	Teaching Assistant The University of Western Ontario 2011-2013 Research Assistant The University of Western Ontario 2010-2011

Publications:

Doyle, Kevin. (2011). *Facet Level Personality Predictors of Perceptions of Group Processes*. Poster presented at the annual meeting of the Canadian Psychological Association, Toronto, ON.