5-2010

The Effect of an Overall Rating Item on Halo Error in Performance Evaluations

S. Elizabeth Hogue
Western Kentucky University, sarah.lawrence624@wku.edu

THE EFFECT OF AN OVERALL RATING ITEM ON HALO ERROR IN PERFORMANCE EVALUATIONS

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

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

By
S. Elizabeth Hogue

May 2010

THE EFFECT OF AN OVERALL RATING ITEM ON HALO ERROR IN PERFORMANCE EVALUATIONS

Date Recommended__March 25, 2010____

__Reagan Brown_____
Director of Thesis

__Elizabeth Shoenfelt_____

__Tony Paquin_____

_____
Dean, Graduate Studies and Research Date

Table of Contents

THE EFFECT OF AN OVERALL RATING ITEM ON HALO ERROR IN PERFORMANCE EVALUATIONS

S. Elizabeth Hogue May 2010 Pages 19

Directed by: Reagan Brown

Department of Psychology Western Kentucky University

This study focuses on how the presence or absence of an overall rating item on a performance evaluation form affects levels of halo error and satisfaction with the form. Participants included undergraduate college students who were randomly assigned to groups receiving a form with or without an overall rating item at the beginning of the form. A satisfaction item was included on both forms. The analyses included a *z*-test for correlations from independent samples to determine the difference between the two evaluation forms and a *t*-test to determine the difference between the satisfaction scores of the two forms. The analyses indicated that the differences between the groups were not significant for levels of halo error or satisfaction.

THE EFFECT OF AN OVERALL RATING ITEM ON HALO ERROR IN PERFORMANCE EVALUATIONS

Ratings made by an outside source (e.g., supervisor, peer, customer) are a common method used for assessment of employee performance (Guion, 1998), and there are many different types of rating systems that can be used. When rating employees some form of criteria must be used to judge performance. Without such a standard a performance rating is meaningless. As regards criteria, there is contention in the literature (e.g., Ghiselli, 1956; Nagle, 1953; Schmidt & Kaplan, 1971) concerning the dimensionality of the criteria used to measure performance. In short, one can measure a unidimensional amalgam of performance (e.g., overall performance, a composite criterion), or one can measure a multidimensional picture of performance (i.e., multiple criteria).

There are many facets to a job; therefore, it has been argued that job performance should be measured in a manner that measures the multiple aspects of a job. Advocates of using multiple criteria promoted the idea that measures of different variables (i.e., different parts of job performance) should not be combined into a single criterion because most jobs are multidimensional and because an employee cannot be described uniquely by combining independent scores (e.g., Ghiselli, 1956; Schmidt & Kaplan, 1971). Therefore, combining scores would lead to a single criterion score that is essentially impossible to interpret (Ghiselli; Schmidt & Kaplan). Especially in the case where different variables have low correlations with each other, the proponents of multiple criteria insisted that these variables are measuring different things and therefore should not be combined (Schmidt & Kaplan).

Conversely, instead of assessing each criterion separately, multiple criteria can be combined to create a composite criterion, which would indicate an employee's overall success (Schmidt & Kaplan, 1971). Proponents of this approach contend that combining scores into a single criterion is necessary to be able to compare individuals for dichotomous decisions in the organization and to determine the value of the employee to the organization as a whole (e.g., Nagle, 1953; Schmidt & Kaplan).

Schmidt and Kaplan (1971) proposed a resolution to the criteria controversy. Their proposal was not so much of a decision for one or the other, but a compromise between the two. The authors asserted that both multiple and composite criteria are applicable depending on the reason the criteria is being used. If the reason is to achieve practical goals such as defining an employee's economic contribution to the organization or to be able to compare employees based on performance then a composite criterion is useful. However, if the goals to be achieved are psychological, such as increased understanding of the different aspects of job performance, then multiple criteria can be used to assess the various aspects of the job.

Despite the appropriateness of the different types of criteria, none of the arguments presented address which type of criteria leads to more accurate measurement of job performance ratings. It is possible that using a composite criterion may lead to more accurate measurement of ratings if there is a general factor present in job performance. A general job factor has been hypothesized because the same traits and abilities are needed to perform multiple dimensions of a job (Viswesvaran, Schmidt, & Ones, 2005).

Viswesvaran et al. (2005) performed a meta-analysis to determine if a general factor is present in job performance ratings. They compiled 90 years of performance rating data and found that a general factor accounted for a large portion of the variance among correlations of job performance dimensions. If a general factor is present, their finding indicates that it would be justifiable to combine criteria into a single composite criterion. It also signifies that an overall measure of job performance would have construct validity and not just be a combination of unrelated measures (Viswesvaran et al.).

The presence of a general factor also raises the issue of combining multiple aspects of job performance to form a composite criterion versus measuring an overall aspect of performance singularly. As mentioned previously, Schmidt and Kaplan (1971) proposed that a composite criterion is acceptable if the use of a composite is consistent with the goals of the evaluation (e.g., summative evaluation, comparison between individuals, determining overall job value to the organization). In contrast, the presence of a general factor indicates the ability to use a single item, as opposed to a composite of many items, to measure an individual's overall performance with no loss of measurement precision.

If a general factor exists, then use of a single overall rating item in performance ratings may be justified. An overall rating item is one that asks about the overall performance of the person being rated. If a general factor in job performance allows for the use of a composite measure, then it is possible that the use of a broad item relating to overall performance is also valid. Cashin and Downey (1992) examined overall rating items on student evaluations of teacher performance. The overall rating items that they

used were “Overall, I rate this INSTRUCTOR an excellent teacher” and “Overall, I rate this an excellent COURSE” (Cashin & Downey, p. 572). The authors found that the two items accounted for the majority of the variance in the student ratings, 54% and 60%, respectively, and concluded that the overall items provide useful information in student ratings when the results are used for summative evaluation or personnel decisions.

There are issues that are inherent in using overall performance ratings that should be addressed if these items are to be included in performance rating instruments. A problem that occurs frequently in performance ratings is halo error. Halo error occurs when a rater’s overall impression of the person being rated influences ratings of specific attributes (Murphy, Jako, & Anhalt, 1993). Even when the rater possesses enough information to make independent ratings of the separate dimensions, halo error causes ratings to be consistent with the rater’s overall impression (Jacobs & Kozlowski, 1985). Halo error is a combination of two parts, true halo and illusory halo. True halo is defined as the actual overlap that occurs between the dimensions being rated whereas illusory halo is true halo plus any irrelevant factors that influence the rating such as measurement errors and systematic rating errors (Murphy et al.). True halo is not error in that it represents actual consistency of performance across multiple dimensions (Murphy et al.).

Murphy et al. (1993) concluded that many of the efforts that have been made to remove halo are unnecessary. The authors suggested that halo error might not be a serious problem at all. Rather, it may actually serve to increase accuracy and utility in ratings especially when the ratings are used to distinguish between persons being rated. Halo error may make it more difficult to discriminate between the individual attributes of a person (e.g., personal strengths and weaknesses), but it may contribute to distinguishing

between ratees on overall levels of performance (Murphy et al.). The authors proposed that administrative decisions are often based on the rank ordering of individuals by an overall score and are unaffected by the variability among the individual scores each person receives. If this is the case, halo error may not be a serious problem.

Because of the supposed pervasiveness of halo error, there have been many attempts at trying to remove it from ratings, especially through statistical control (Murphy et al., 1993). The authors concluded that these techniques (e.g., partialing out the overall rating from the separate dimensional ratings) do not function as desired because they seem to remove illusory halo but also true halo as well.

Murphy et al. (1993) also contended that controlling halo is not necessary because halo can actually increase the reliability of ratings by increasing the correlations between items. This argument is flawed because the increase in the correlations among the dimensions, if even partly due to illusory halo, is artifactual due to the fact that illusory halo is error. That is, an increase in correlation due to error does not result in beneficial outcomes. Thus, the reliability increase associated with halo error is not truly higher, only apparently higher. As an argument by analogy, consider a test-retest reliability study involving a group of people completing a self-report test of extraversion. If 25% of the test takers exhibited a leniency, severity, or central tendency bias at both administrations, the correlation between scores from the two administrations would be inflated, resulting in an overestimate of the test's reliability. Would one argue that having a sample of test takers composed of at least some people with obvious rating biases is a good thing? No, the rating biases are a source of error in the study. It just happens to be the case that these particular errors result in an increased reliability estimate.

If illusory halo could be removed while retaining true halo, controlling for this erroneous part of halo would be desirable. However, ratings are composed of both true and illusory halo, and the distinction between them, in essence, is theoretical (Murphy et al., 1993). In a field setting, it is impossible to separate true halo (i.e., accurate ratings of the relationships among dimensions) from illusory halo (e.g., errors in judgment and observation, memory failures, rater tendencies, etc). Thus, in field studies it is also impossible to determine the amount of true halo that is actually present in the ratings (Murphy et al.). Consequently, rather than attempt to remove illusory halo, Murphy et al. suggested that researchers should perform studies that help identify situations in which halo may or may not be a problem. From these studies, the information gained regarding halo could then be used in a practical setting.

The purpose of the current study is to discover the effects of an overall rating item on halo error in performance evaluation ratings. This research will investigate whether an overall item included in a performance rating affects the amount of halo observed. The study will include two groups completing performance evaluation forms, one form including an overall rating item and one without. Because the only difference between conditions will be the presence or absence of an overall item, any changes in halo will be due to illusory halo (i.e., halo error, not halo effect). Murphy et al. (1993) proposed that if a rater is given an overall item at the beginning of an evaluation, halo error may increase due to the fact that the rater's attention has been drawn to making judgments of the ratee's overall performance. However, the persistent presence of halo error also suggests that it is possible raters may tend to think in terms of generalities regarding the ratee's performance. If the latter scenario is the case, it could be the case that allowing a rater to

give an overall rating at the beginning of the evaluation may decrease halo error by satisfying the rater's desire to make an overall evaluation, freeing up cognitive attention to focus on the specific dimensions that follow. Also, because raters do tend to make general evaluations of ratees, raters may want the opportunity to make an overall rating. Therefore, including an overall rating item may also increase the rater's satisfaction with the instrument.

Hypothesis 1: Ratings obtained using performance appraisal forms with an overall rating item at the beginning of the form will differ in levels of halo error from ratings obtained using forms without an overall item.

Hypothesis 2: Raters will be more satisfied with appraisal forms which include an overall rating item as compared to those forms without an overall item.

Method

Participants

Two hundred and forty-three participants were recruited from undergraduate classes in a variety of fields (Appendix A) at a southeastern university. Consent was obtained from the course professors to use their students as raters and from the students prior to their participation.

Materials

Two forms of a teacher performance evaluation were used. The forms were identical except for the inclusion of an overall rating item at the beginning of one form (Appendix B) and the absence of the overall item on the second form (Appendix C). A satisfaction item was included at the end of both versions of the evaluation forms to assess participant satisfaction with the rating process. The rating options on the forms ranged from Strongly Agree (1) to Strongly Disagree (5).

Design

Correlations between items or dimensions are typically employed to index the degree of halo error, with moderately high or higher correlations interpreted as indicating the presence of halo (Jacobs & Kozlowski, 1985). The present study was designed to test the difference between two rating forms; correlations from independent samples (one sample per form) were compared to determine which form had more halo error (i.e., illusory halo). The difference between the raters' satisfaction was tested to determine if the mean satisfaction scores differed between the two versions of the form. Participants were randomly assigned to either the overall rating condition or the control condition (i.e., no overall rating) by alternating forms during distribution within a class. Some

participants were excluded from the data analyses due to incomplete data. For the correlational analysis, the final participant total in each group was 112 participants in the overall rating group and 111 participants in the control group. For the satisfaction analysis, the final participant total was 120 participants in the overall rating group and 122 participants in the control group. Differences in sample sizes between the two analyses (specifically, the loss of approximately 10 participants from each group for the correlational analyses) were due to the presence of non-responses to some items.

Procedure

Instructors were asked to allow their class the opportunity to participate in the research. Instructors and students were informed that the study was the thesis project of a university graduate student, was confidential, and had no bearing on the professor's official performance ratings for the university. Standard instructions (Appendix D) were read to the participants in every class. At the conclusion of the study, all participants and professors were thanked for their time and participation.

Analyses

Coefficient alphas were computed to assess the average correlation among items on the evaluation forms. A *z*-test for correlations from independent samples was used to determine if there was a significant difference between the coefficient alphas. Random assignment to conditions within each classroom potentially eliminated true halo as a source for the difference between conditions. Satisfaction was measured by examining the mean satisfaction scores of the two groups. To assess the difference between raters' satisfaction scores, a *t*-test was used to determine if there was a significant difference between the mean satisfaction scores of the two forms. Despite the differences in class

type (i.e., different classes were taught by different instructors with differing levels of quality), the analyses were not executed separately within class but with data pooled across classes. Pooling across classes allowed for a larger sample size, which increased power. In order to determine if combining the classes would be adversely affect the study, a Monte Carlo analysis was completed. The results of the analysis indicated that pooling data across classes would not confound the results as long as participants within each class were randomly assigned to groups and group sizes were equal within each class.

Results

The first hypothesis, which stated that internal consistency will differ by test type, was examined by a comparison of coefficient alphas as well as split half (odd/even split) correlations. Coefficient alpha for the overall rating condition was .928 (n = 112) and was .918 for the no overall rating condition (n = 111). A z-test for correlations from independent samples indicated that the difference between the overall rating and the no overall rating conditions was not significant, z = .52, p > .05. A comparison of split-half reliability coefficients (odd items vs. even items) yielded similar results, r = .887 for overall rating condition, r = .870 for no overall rating condition, with a non-significant difference between conditions, z = .55, p > .05.

The second hypothesis, which stated that raters in the overall item condition would be more satisfied with the rating process, was analyzed with a t-test. The mean satisfaction scores were 1.32 (SD = .502, n = 120) for the overall rating condition and 1.39 (SD = .537, n = 122) for the no overall rating condition. The rating scale ranged from 1 (*strongly agree*) to 5 (*strongly disagree*). The t-test indicated that the difference between the overall rating condition and the no overall rating condition was not significant, t (240) = -1.03, p = .306.

Discussion

Halo error occurs when a rater's overall impression of the person being rated influences that rater's ratings of specific attributes (Murphy et al., 1993). Observed correlations among ratings of specific components of job performance are combinations of true and illusory halo. True halo is defined as the actual overlap between the dimensions being rated whereas illusory halo is halo caused by any irrelevant, erroneous factors that influence the rating (Murphy et al.). True halo cannot be separated from illusory halo in field studies (Murphy et al.); therefore, the current study used random assignment to two conditions (i.e., receiving an evaluation form with or without an overall rating item) to control for true halo as the source of the difference between groups. Any significant difference would be due to illusory halo. Thus, the purposes of the current study were to investigate the effects of an overall rating item on halo error in performance evaluation ratings and to determine if an overall item included in a performance rating affected the amount of halo observed.

Another purpose of this study was to determine if raters are more satisfied when an overall item is included on the rating form. Both forms included a satisfaction item. A *z*-test was used to determine if a difference existed between the coefficient alpha of each group, and a *t*-test was used to determine if there was a difference between the mean satisfaction score of each group.

The data did not support the hypotheses of the study. The results suggest that there is no difference in the amount of halo error in rating forms with or without an overall item. Consequently, the presence of an overall item does not appear to affect the amount of halo error in performance ratings. In a practical setting this could suggest that

raters are likely to give the same type of ratings regardless of the presence of an overall item at the beginning of the form. Thus, although this study does not reveal ways to reduce halo error, it does suggest that an overall item will not increase halo error.

There was also no difference between the satisfaction of raters with a form that included or excluded an overall rating item. The results indicate that the desire to make an overall rating may not affect a rater's satisfaction with a form even if that option is not offered to the rater. Alternatively, it may be the case that very few raters have a desire to make an overall rating. On a practical level, this suggests that the inclusion of an overall item is not necessary for raters to feel satisfied with a rating form.

One limitation to the present study is that the study was performed with students who knew that the ratings were for graduate thesis work and not official ratings. This knowledge may have caused some students to refrain from making the effort necessary to provide accurate ratings.

One possible reason for the lack of differences between the conditions relates to the amount of halo error (due to true and illusory halo) already present in the ratings. The coefficient alphas for both conditions were greater than .90; therefore, given an upper limit of 1.0, there was little opportunity (via the presence of an overall rating) for additional halo error to be expressed. A study by Feeley (2002) indicated that halo is greater in conditions when raters are unmotivated and/or unconcerned about the results of the evaluation. Given this finding, it is not surprising that raters who were unmotivated or unconcerned (i.e., instructed that the ratings did not count for any applied purpose) would provide highly haloed ratings in both conditions.

Similarly, it is worth noting that raters using a form without an overall rating item were highly satisfied (i.e., more than half strongly agreed with a statement that said "I am satisfied with the items on this form"). Such a high level of satisfaction with the form that did not include an overall rating item leaves little room for differences between the two conditions.

In summary, the inclusion or exclusion of overall rating items at the beginning of a rating form does not appear to affect the magnitude of halo error or the satisfaction of the raters with the form itself. Despite the outcome, future research on this topic could be beneficial. Studies performed with larger sample sizes and in a setting where the results of the performance ratings yield tangible outcomes such as raises or promotions would be valuable to determine if overall rating items can have an effect on halo error. Because halo error is common (Jacobs & Kozlowski, 1985; Murphy et al., 1993), additional research that further explores the understanding of what increases or decreases the magnitude of halo error would be helpful.

References

Cashin, W. E., & Downey, R. G. (1992). Using global student rating items for summative evaluation. *Journal of Educational Psychology, 84,* 563-572.

Feeley, T. H. (2002). Evidence of halo error in student evaluations of communication instruction. *Communication Education, 51,* 225-236.

Ghiselli, E. E. (1956). Dimensional problems of criteria. *Journal of Applied Psychology, 40,* 1-4.

Guion, R. M. (1998). *Assessment, measurement, and prediction for personnel decisions.* Mahwah, NJ: Lawrence Erlbaum Associates.

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

Murphy, K. R., Jako, R. A., & Anhalt, R. L. (1993). Nature and consequences of halo error: A critical analysis. *Journal of Applied Psychology, 78,* 218-225.

Nagle, B. F. (1953). Criterion development. *Personnel Psychology, 6,* 271-289.

Schmidt, F. L., & Kaplan, L. B. (1971). Composite vs. multiple criteria: A review and resolution of the controversy. *Personnel Psychology, 24,* 419-434.

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology, 90,* 108-131.

Appendix A

Courses Used in Data Collection

1. Counseling Children and Adolescents
2. Techniques of Counseling
3. Testing and Assessment
4. Introduction to Psychology
5. Pain Control in Dentistry
6. Using Statistics in Sociology
7. Communication Disorders
8. Social and Cultural Diversity
9. Introduction to Counseling
10. Classroom Behavior
11. Teaching Strategy for Secondary Schools
12. Techniques of Counseling
13. Introduction to Teacher Education
14. Statistics/Experimental Psychology
15. Educational Psychology

Appendix B

Rating Form with Overall Item Included

Please read each item carefully and answer by placing an ***X*** in the appropriate box.
Use the following scale to indicate your answer to the items below:

Strongly Agree (**SA**) Agree (**A**) Neutral (**N**) Disagree (**D**) Strongly Disagree (**SD**)

Item	SA	A	N	D	SD
Overall, my instructor is effective.					

The following items describe specific aspects of performance. Each item should be rated independently/separately from the other items. For example, if a secretary is good at typing but bad at scheduling appointments, you would strongly agree that the secretary is a good typist and strongly disagree that the secretary is good at scheduling appointments.

Item	SA	A	N	D	SD
My instructor is helpful with regard to answering questions or explaining concepts outside of class.					
My instructor arrives to and starts class on time.					
My instructor is prepared for all lectures and class activities.					
My instructor displays interest in teaching this class.					
My instructor treats me fairly with regard to race, age, sex, religion, national origin, disability, and sexual orientation.					
My instructor is able to explain course material in a clear, concise manner.					
My instructor displays a clear understanding of course topics.					
Grades are fairly assigned.					
The assignments are helpful in learning the material in this course.					
The tests accurately measure how much I learned in this course.					
My instructor interacts with students in a personable manner.					
My instructor returns graded materials in a timely fashion.					
My instructor chooses materials (textbooks, articles, etc.) that are helpful in learning the course topics.					

Finally, please rate how well this rating form allowed you to express your opinion of the instructor.

Item	SA	A	N	D	SD
I am satisfied with the items on this form.					

Appendix C

Rating Form without Overall Item

Please read each item carefully and answer by placing an ***X*** in the appropriate box.
Use the following scale to indicate your answer to the items below:

Strongly Agree (**SA**) Agree (**A**) Neutral (**N**) Disagree (**D**) Strongly Disagree (**SD**)

The following items describe specific aspects of performance. Each item should be rated independently/separately from the other items. For example, if a secretary is good at typing but bad at scheduling appointments, you would strongly agree that the secretary is a good typist and strongly disagree that the secretary is good at scheduling appointments.

Item	SA	A	N	D	SD
My instructor is helpful with regard to answering questions or explaining concepts outside of class.					
My instructor arrives to and starts class on time.					
My instructor is prepared for all lectures and class activities.					
My instructor displays interest in teaching this class.					
My instructor treats me fairly with regard to race, age, sex, religion, national origin, disability, and sexual orientation.					
My instructor is able to explain course material in a clear, concise manner.					
My instructor displays a clear understanding of course topics.					
Grades are fairly assigned.					
The assignments are helpful in learning the material in this course.					
The tests accurately measure how much I learned in this course.					
My instructor interacts with students in a personable manner.					
My instructor returns graded materials in a timely fashion.					
My instructor chooses materials (textbooks, articles, etc.) that are helpful in learning the course topics.					

Finally, please rate how well this rating form allowed you to express your opinion of the instructor.

Item	SA	A	N	D	SD
I am satisfied with the items on this form.					

Appendix D

Instructions

I'm doing research for my graduate thesis in Industrial/Organizational Psychology. Your professor has offered me your assistance. You will be filling out a short questionnaire about your professor's performance in this course. This is completely voluntary and you can choose not to participate at any time. All of your answers will be confidential. In fact, please don't put your name or any type of identification, like your student ID numbers, on the questionnaires. These ratings are strictly for research purposes and have no effect on your professor's official performance ratings for [university name]. First I am going to hand out an informed consent form that simply says that you agree to participate in my research, that it will not cause you any harm, and that you may stop participating at any time you choose. I need you to sign and date these.

Now I will hand out the questionnaire. On the questionnaire you will read items about your professor's performance and rate them on how much you agree or disagree with the statement. The scale for the ratings ranges from Strongly Agree to Strongly Disagree. Please read all the directions and answer all the items. If you have any questions, please feel free to ask.