


2004

## Bias in the Evaluation Process: Influences of Speaker Order, Speaker Quality, and Gender on Rater Error in the Performance Based Course

Paul D. Turman  
*University of Northern Iowa*

Matthew H. Barton  
*Southern Utah University*

Follow this and additional works at: <http://ecommons.udayton.edu/bcca>

 Part of the [Higher Education Commons](#), [Interpersonal and Small Group Communication Commons](#), [Mass Communication Commons](#), [Other Communication Commons](#), and the [Speech and Rhetorical Studies Commons](#)

---

### Recommended Citation

Turman, Paul D. and Barton, Matthew H. (2004) "Bias in the Evaluation Process: Influences of Speaker Order, Speaker Quality, and Gender on Rater Error in the Performance Based Course," *Basic Communication Course Annual*: Vol. 16 , Article 6.  
Available at: <http://ecommons.udayton.edu/bcca/vol16/iss1/6>

This Article is brought to you for free and open access by the Department of Communication at eCommons. It has been accepted for inclusion in Basic Communication Course Annual by an authorized editor of eCommons. For more information, please contact [frice1@udayton.edu](mailto:frice1@udayton.edu), [mschlangen1@udayton.edu](mailto:mschlangen1@udayton.edu).

## **Bias in the Evaluation Process: Influences of Speaker Order, Speaker Quality, and Gender on Rater Error in the Performance Based Course**

---

*Paul D. Turman  
Matthew H. Barton*

Demand for increased proficiency in communication skills has increased dramatically in recent years (Sawyer & Behnke, 1997). Consequently, the basic course has taken the brunt of this demand. Current trends in higher education demonstrate that the basic course at most universities will find itself servicing even more students in the near future. According to the National Center for Educational Statistics, the number of high school students continuing on with their education after graduation increased by 12% between 1995 and 2002, and as a result college enrollment has increased by 17% in this same time period (public and private not-for-profit institutions). If higher education continues to see a persistent influx of students in the wake of current economic conditions, the increasing student population will begin to place a significant burden on current basic course structures.

Increasing the number of sections offered in the basic course has been the traditional solution to the problem of increased demand (Gibson, Hann, Smythe, & Hayes, 1980; Gibson, Hanna, & Huddleston, 1985; Sawyer & Behnke, 1997). However, this strategy comes with

a number of pitfalls. First, the buildup of additional sections requires an increase in the size of the instructional staff. This move is difficult to justify with so many demands on already strained departmental and institutional budgets (Fedler & Smith, 1992). Second, when the addition of staff is warranted, administrators often provide increases in personnel in the form of adjunct or part-time faculty, which provide only temporary solutions for most basic course directors (Sawyer & Behnke, 1997). On the other hand, some departments, particularly those at larger institutions, have increased the utilization of graduate teaching assistants (Buerkel-Rothfuss & Gray, 1990; Roach, 1991; Williams & Roach, 1993; Williams & Schaller, 1994). While this action has reduced some of the pressure, it seems that administrators are “upping the ante” by adding more and more students to these courses. Thus, instead of solving the problems associated with increased class size, they are perpetuated. Moreover, in their assessment of the basic course, Gibson, Hanna & Huddleston (1985) found that a majority of colleges and universities utilized either a public speaking (54%) or a hybrid (34%) course structure suggesting that the basic course continues to place an emphasis on student performance.

Research has identified three primary problems that need to be addressed. First, although increasing the number of sections available for the basic course is one available option, increasing class size places significant restrictions and limitations on the function of a performance based course and ultimately limits students’ ability to obtain communication competence (O’Hair, Friedrich, Wiemann, & Wiemann, 1995). Second, larger class sizes pose a number of pragmatic problems that

need to be addressed (Cheatham & Jordan, 1972). For instance, in order to provide larger classes of students with the opportunity to practice and receive feedback on speeches, instructors are forced to either add more speech days or add more speakers on a given day. In some cases they must do both. Instructors who have taught performance-based courses have likely had groups of three or four speech days throughout the semester where they have heard as many as eight or more speakers on each of those days, which can contribute to the potential for rater fatigue. This predicament is compounded by the fact that many instructors teach more than one section of the basic course, meaning that they may encounter 16 to 24 speakers on each of those days. Considering the other responsibilities of faculty life, instructors want and need to be more efficient. Rater error can happen not because instructors are unconcerned about improving student speaking skills, rather because they have limited time to grade presentations in detail with so many speakers to evaluate. Thus, cutting corners in the evaluation process becomes a greater temptation. Finally, hearing so many speeches over a consistent time decreases the odds that meaningful distinctions between speakers can be consistently accomplished (Miller, 1964). Consequently, the purpose of this study is to examine if a potential *evaluation threshold* exists in the basic communication course (e.g., those with a strong public speaking or performance-based component). Logic and experience suggest that there may be a limited number of student speeches that can be effectively evaluated in a given class period without compromising the quality and quantity of instructor feedback. Specifically, this study attempts to examine

situational qualities (e.g., presentation quality and speaker order), which may further contribute to grading inconsistencies.

## REVIEW OF LITERATURE

To be successful in higher education, communication faculty must learn to provide effective feedback that is detailed, individualized, consistent and objective (Bock & Bock, 1981). Reaching this level of success is obviously a difficult undertaking because of a number of factors. For an instructor to arrive at a score or final grade for a presentation, he/she is required to assess the quality of that performance. The expectation is that the best presenter will receive the highest score regardless of the individual rating of the presentation (Lunz, Wright, & Linacre, 1990). Saal, Downey and Lahey (1980) indicated that although the expectation for unbiased scoring is connected with the performance appraisal process, research examining the subjectivity associated with rater error has identified significant variations regardless of the type of appraisal (e.g. job performance, leadership evaluation, personnel selection, etc.). Engelhard (1994) argued that one of the major problems with appraisal processes is that they depend primarily on the quality of experts who make the final judgment. In one of the first examinations of rater error, Guilford (1936) stated that "Raters are human and they are therefore subject to all the errors to which human-kind must plead guilty" (p. 272). When rater error does occur it has the potential of weakening the reliability and validity of the system employing the assessment,

and information provided by the assessment (Bannister, et al., 1987). Evaluations of rater validity and reliability have reported coefficient levels ranging from .33 to .91 (Dunbar, Kortez, and Hoover, 1991) and .50 to .93 (Vand Der Vleuten & Swanson, 1990) which suggests that as the range of error increases the potential for accurate assessment will decline significantly.

As the preceding studies have indicated, the existence of rater error is a legitimate problem when subjective assessment is involved. Also, depending on the situation facing the rater, error can be a result of a number of factors including: the assessment tool used, the scoring procedures, and individual rater bias (Popham, 2002). First, the flaws in assessment tools can be caused by a deficiency in the evaluation criteria being used. As a result inappropriate ratings are made because of the ambiguity associated with the methods used to score certain behaviors described in the evaluation criteria (e.g., one instructor may view eye contact while another may look for gestures as the most important part of the delivery). Second, ambiguity or flaws in the scoring procedures occur when raters are asked to assess too many qualities about a particular ratee (Popham, 2002).

The third and perhaps most significant type of assessment error is a result of bias within the individual rater. Individual rater error has seen significant research in the past century and this body of literature has identified three primary types of errors that occur at the individual level. The most prominent is the *halo effect* first identified by Thorndike (1920) during the examination of consistency across evaluations for officer candidates in the military. When applied to an educa-

tional context, Engelhard (1994) suggested that the halo effect would occur when a teacher's impression or previous experience with a particular student affected the score obtained on the assessment. As a result, the halo effect can occur in one of two ways; if the impression is favorable the rating will be higher, and if it is unfavorable the rating will be lower. The halo effect has also been attributed to a rater's unwillingness to make distinctions across various dimensions on a rating scale and as a result they place ratees at the same level across all criteria dimensions. Although research applying the halo effect to student presentations has been limited, Harper and Hughey (1986) identified literature demonstrating that instructors "receive more favorably the communication performances of students who possess similar communication attributes" to their own (p. 147).

Another individual rater error that has been identified is called *positive leniency/rater severity* (Engelhard, 1994), where the rater has a tendency to consistently provide ratings on either the high or low end of the scale, making their assessment practices unfair. Positive and negative leniency can also be a function of attribution error on the part of the rater. These types of errors occur more at the holistic level, when instructors are more likely to grade all students higher than they should, or the converse happens when they choose to be more critical of all student behaviors than is logically warranted.

Finally, *central tendency* or *restriction of range* occurs when ratings are "clustered around the midpoint of the rating scale, reflecting rater reluctance to use either of the extreme ends of the continuum" (Saal, Downey, &

Lahey, 1980, p. 418). This type of individual rater error reflects how the rater utilizes the categories on the rating scale itself. Engelhard (1994) suggested this type of error is most likely to occur when raters use the evaluation criteria differently by which some overuse extreme categories and others overuse those categories in the middle of the scale.

Research specific to rater error in the context of speech assessment is relatively limited to date, however previous communication research has suggested a need to be concerned with primacy and recency effects during the assessment process. For example, in 1925, Lund explored a theory that he called primacy, which referred to the notion that an idea presented first in a discussion would have a greater impact than the opposing side presented second (in Mason, 1976). Other research has since followed Lund's lead exploring the viability of his theory (Anderson & Barrios, 1961; Bishop, 1987; Ehrensberger, 1945; Freebody & Anderson, 1986; Jersild, 1929; Krosnick & Alwin, 1987; Sato, 1990). Specifically relating to public speaking, Knowler (1936) found that competitive speakers in first and last positions are more commonly ranked in intermediate positions as opposed to either high or low extremes and second to last speakers often score highest on final averages. Benson and Maitlen (1975) disputed some of Knowler's findings as their research concluded that there was no significant relationship between rank and speaking position. To test the effectiveness of the Instructor Assistant training process and grading procedures Turman and Barton (2003) explored primacy and recency effects as a result of speaker order. Four groups of undergraduate raters were asked to grade four ten-minute persuasive



speeches after participating in an extensive training program. Presentations were placed in varying orders for each group and no evidence of primacy or recency influence or rater error emerged across groups, indicating speaker order had no impact on the final grades students received. Aside from this particular study, literature on primacy and recency effects and rater error does not deal directly with speaking situations and it appears to be badly dated (Ehrensberger, 1945; Lund, 1925 in Mason). Ironically enough however, there are findings favoring both types of effects (Krosnick & Alwin, 1987; Miller & Campbell, 1959).

### ***Research Questions***

Research on general rater error (halo effect, severity and leniency, and central tendency) has suggested that the subjectivity associated with evaluation of human performance guarantees the potential for error in performance appraisal. However, research on rater error in the context of communication and speech performance has presented inconclusive results when examining the influence of rater error on speaker order. Additionally, these findings do not indicate whether rater error is unlikely to exist in situations where more than four speakers are evaluated in a given class period (Turman & Barton, 2003). Also, research has yet to represent a design which is reflective of a typical speech day (e.g. grading student speeches of varying quality) which might increase the potential for rater error. In other words, when examining what occurs in a traditional classroom structure one would expect to find seven or eight students speaking on a given day coupled with variations in the speaking order and in the quality of

student speeches, resulting in a likely variability in student scores related to these factors. Thus, to isolate and clarify the potential influence of speaker order and quality when the number of speakers is increased, the following research question was set forth.

RQ1: Does speaker order and presentation quality influence the subsequent grade that students receive?

An additional challenge raters face is providing effective feedback to students, while ensuring that their grading practices are both valid and reliable. One of the primary objectives of a course with a presentation focus is to provide students with effective feedback to enhance their speaking ability over the course of a semester (O'Hair, Friedrich, Wiemann, and Wiemann, 1995; Sawyer & Behnke, 1997). Because of the ego involvement associated with public speaking situations, feedback providing more than a simple numerical justification for student grades is necessary. Raters are expected to provide students with high quality feedback by which students engage in skill building as a way to become stronger public speakers. One could argue that in addition to increased potential for rater error based on speaker order, raters may also experience rater fatigue, and consequently be less likely to provide high quality feedback as they progress through the speaker order. While proving fatigue is difficult, the present study is concerned with finding any hint of fatigue that may influence the evaluation process and provide an additional avenue of research in the context of rater error. Overall, the assumption of the following research question implies that students presenting presentations at the

beginning of the speaker order would receive higher quality comments than those at the end, suggesting that fatigue is present and may account for this discrepancy. To analyze the potential for this assumption, the following research question was set forth:

RQ2: Does the order in which a speaker presents influence the quality of comments and feedback provided by the rater?

In addition to the preceding problems, limited research has attempted to determine the influence of other mediating variables on rater error. For example some studies have explored the problems associated with the way that international students (Young, 1998) and students with different dialects (Agee & Smith, 1974) are evaluated. However, a more obvious influence on rater error comes from an examination of gender. Exploration into gender as a significant problem related to speech evaluation has found that women tend to be more lenient graders than men when using rating scales (Bock, 1970), drawing attention to the need for adequate assessment tools. In addition, Bock and Bock (1977) found that instructors demonstrated a tendency to rate students of the same sex more highly, commonly known as a *trait error*, which occurs when instructors place an over-emphasis on a specific trait or skill (Ford, Puckett & Tucker, 1987; King, 1998). Thus, there appears to be a precedent set for a negative evaluation bias based on gender that needs to be addressed more completely. In an attempt to determine whether the gender of the rater influenced student grades based on the speaker's gender, the following research question was set forth:

RQ3: Does rater gender influence the quality of comments students receive for classroom presentations?

## METHOD

### *Participants & Procedures*

*Raters.* The raters in this study consisted of 76 (males,  $n = 30$ ; females,  $n = 46$ ) undergraduate students currently working with the basic course at a large Midwestern university. Raters were competitively selected from a pool of students who had successfully completed the basic course by utilizing grade point average and reported performance in the classroom. Raters were given course credit for their participation and included a mixture of students from a variety of majors (e.g., communication studies, business, etc.).

*Training Procedures.* To prepare for the assessment process raters were required to complete an eight-week training program which focused on evaluation of recorded presentations and speaker outlines. Before grading any of the presentations, the primary researchers familiarized the raters with a criterion referenced evaluation instrument which was divided into three major sections (i.e., introduction and conclusion, body, and delivery). Over the course of the eight week training period, the raters were trained to utilize the evaluation form which assigned specific point values to respective elements for each of the three major criteria sections. Twenty points were assigned to the introduction and conclusion (e.g., assessment of things such as the

attention getter, preview and summary statements, and closing remarks), 40 points reflecting content (e.g., main point development, organizational structure, documentation and use of evidence), and 40 points for delivery (e.g., including eye contact, extemporaneous delivery style, gestures, posture, and movement). Additionally, grading techniques such as taking copious notes, utilizing positive and negative comments, and the need for providing appropriate feedback were addressed to further ensure consistency across rater use of the evaluation form. Each reviewer viewed and assessed ten presentations, entered into discussion with fellow reviewers concerning the comments and grades assigned, and then submitted their evaluation forms for assessment by the primary researchers.

### ***Experimental Design***

To obtain a pool of student presentations, 25 speeches were taped from one section of the basic course for a persuasive speech assignment. The primary researchers each evaluated the presentations and assigned grades based on the same criterion referenced evaluation instrument (intercoder reliability was calculated at .89). From these presentations, the primary researchers utilized a cluster sampling technique to select two speeches from each of the *A*, *B*, *C*, and *D* grade categories ( $n = 8$ ). Also, to incorporate gender as an independent variable, male ( $n = 4$ ) and female ( $n = 4$ ) students were selected at each grade category as well. Those speeches selected for utilization in this study ranged in length from 7 to 9 minutes, and after the selection process, presentations were re-taped in varying order utilizing an incomplete factorial design (see

Table 1 for representation of the distribution of multiple A through D presentations across the treatment groups)<sup>1</sup>. Additionally, thirty-second delays were incorporated into each tape between each speaker to simulate the amount of time graders often utilize between speakers on a typical presentation day in the classroom.

Table 1  
Speaker Order Assignments for Treatment Groups

Speaker Position	Rater Groups							
	1	2	3	4	5	6	7	8
1 <sup>st</sup>	A-1	D-2	D-1	C-1	A-2	D-2	B-1	C-2
2 <sup>nd</sup>	A-2	D-1	C-1	C-2	B-2	C-2	B-2	C-1
3 <sup>rd</sup>	B-1	B-2	C-2	D-1	C-1	B-2	D-1	A-2
4 <sup>th</sup>	B-2	B-1	A-1	D-2	D-1	A-1	D-2	A-1
5 <sup>th</sup>	C-1	A-1	A-2	A-1	A-1	A-2	C-1	D-1
6 <sup>th</sup>	C-2	A-2	B-1	A-2	B-1	B-1	C-2	B-1
7 <sup>th</sup>	D-1	C-1	D-2	B-2	C-2	C-1	A-1	D-2
8 <sup>th</sup>	D-2	C-2	B-2	B-1	D-2	D-1	A-2	B-2

To assess the presentations the raters were randomly assigned to one of eight treatment groups. Assistants were used to help administer the study, and each was provided with a detailed list of instructions in

<sup>1</sup> A complete experimental design would have required an additional 56 groups to achieve the total number of possible speaker combinations; and would have required approximately 500 additional raters. Additionally, access to student raters and consistent training personnel was limited to a one-year period based on the existing structure of the basic course at this institution.

order to make sure each group followed the same procedures and had the same experience. Participants were asked to watch all eight speeches, evaluate them, and make the necessary comments. To further represent a typical speech day, the raters were given a 24 hour period to make needed comments and were then instructed to return the evaluation forms to the primary researchers to simulate the actual experience of returning scores to the students. To help maximize external validity and eliminate the potential for confounding variables, the research was conducted in classrooms used during the training session. Also, raters were provided with the same environment, visual equipment and tape quality to help ensure a similar experience across each group. Furthermore, raters were not provided with information concerning the nature and purpose of the study to eliminate the increased potential for a halo effect to emerge.

### ***Scales of Measurement***

*Analytic Grading Form.* Raters used an evaluation instrument that utilizes an analytic method by which content and delivery elements were rated and then summed to generate the final score for the presentation, rather than a holistic approach (using personal judgment when determining the importance of specific traits toward the overall product). In an attempt to determine the effectiveness of each approach, Goulden (1994) found that neither the analytic nor holistic method was more effective at producing a reliable assessment of student presentations. To test the effectiveness of the rater training and evaluation procedures, an initial pilot test was conducted using four persuasive presentations

of similar quality. The speaker order was manipulated and 38 undergraduate raters were assigned to one of four treatment groups. An analysis of variance indicated no significant differences across groups ( $F(3, 124) = .492, p > .05$ ) based on rater evaluations when only four presentations were utilized.

*Evaluation Quality.* Two student coders were selected and asked to evaluate rater comments for each of the presentations based on a semantic differential type scale adapted from an instrument developed by Osgood, Suci, & Tannenbaum (1957). This 12-item scale was created to analyze the quality of student comments based on a combination of the introduction/conclusion, body and delivery. Coders were given the stimulus statement, "What is the quality of the written feedback provided by the evaluator for this presentation" and used a 5-point scale to capture perceptions to the degree that each section (e. g., introduction, conclusion, body, delivery) was: good-bad, valuable-worthless, qualified-unqualified and reliable-unreliable. Inter-coder reliability was calculated at .88 for the two coders.

### ***Data Analysis***

Research question one used an  $8 \times 8$  factorial design to measure the potential change in student presentation grades. The order of the presentations (either going 1<sup>st</sup>, 2<sup>nd</sup>, 3<sup>rd</sup>, 4<sup>th</sup>, 5<sup>th</sup>, 6<sup>th</sup>, 7<sup>th</sup>, or 8<sup>th</sup>) and rater group assignments (group 1, 2, 3, 4, 5, 6, 7, or 8) both served as ran-



dom factors<sup>2</sup>. An analysis of variance (ANOVA) with follow-up analyses using the LSD procedure ( $p = .05$ ) was performed to examine the effects of speaker order and presentation quality on students' grades. An ANOVA was also utilized to analyze data for research question two to determine the influence of speaker order on the quality of comments provided for students. Furthermore, data for research question three was assessed using an independent sample t-test to determine significant differences based on rater gender.

## RESULTS

The first research question inquired whether student ratings would be influenced by speaker placement. ANOVA analysis indicated a significant interaction effect based on rater grouping and presentation score ( $F(7, 49) = 8.88, p < .0001, \eta^2 = .35$ ) and post hoc analysis indicated significant differences across groups for each of the eight presentations. Two particular patterns emerged when examining the differences across groups.

First, a number of speaker positions caused a significant decrease in presentation ratings (See table 2). Specifically, scores on presentation *A-1* and *A-2* declined when preceded by lower quality presentations (see group 5, 6, 7, and 8 in Table 1). The grades assigned to

---

<sup>2</sup> Speaker order and grade quality both served as random factors as a function of the incomplete experimental design utilized for data analysis. Because it was not possible to design a complete experiment incorporating the 64 treatment groups necessary, the primary researchers were forced to randomly assign speaker order and grade quality across the eight groups in an attempt to make inferences across the 64 groups required in a complete design.



each of these speakers appeared to be most affected by speaker order wherein presentation *A-1* ranged in scoring from a high of 84.70 ( $SD = 5.69$ ) to a low of 55.55 ( $SD = 10.82$ ). A clear interaction effect emerged when examining the profile plots for the *A-1* presentation when compared with *D-2* (see Figure 1). In this instance the placement of presentation *A-1* in groups 6, 7, and 8 produced a steady decrease in rater scoring, while presentation *D-2* experienced a significant increase in rater scoring for group 5, 6, and 8. Presentation *A-2* experienced similar variability with raters scoring this presentation high ( $M = 85.44$ ,  $SD = 5.70$ ) while other raters influenced by speaker position and preceding speaker quality rated the presentation significantly lower ( $M = 50.90$ ,  $SD = 14.39$ ). Similar declines in scoring were recorded for presentation *C-1* and *C-2*, whereas scores tended to be affected by placement in close proximity to lower quality presentations (see group 6, 7 and 8 in Table 1).

Second, a number of speaker positions resulted in significant increases in presentation ratings (see Table 2). Scores on presentation *C-1* increased significantly when placed in the beginning or end of the presentation rotation (See group 7 on Table 1). *C-1* experienced a significant decline when placed at the front of the order and followed by lower quality presentations (see Figure 2). Finally, *D* presentations tended to increase significantly when there was significant variability in the speaker order (see groups 5, 6, and 8 on Table 1).

No significant differences, however, were found for research question two which asked whether speaker order would impact the quality of written comments. The ANOVA analysis indicated no significant differences ( $F$





(7,600) = .086,  $p > .05$ ) indicating that those students who present in the last speaking position received the same quality comments as those who present in the first. Research question three assessed whether rater gender would affect the quality of written comments provided to students on the analytic evaluation form. Findings from the T-test indicated significant differences did exist ( $t = (606) = 7.06$ ,  $p = .008$ ), suggesting that female raters provided higher quality written comments ( $M = 14.60$ ;  $SD = 4.43$ ) when compared to male raters ( $M = 15.20$ ;  $SD = 3.79$ ).

## DISCUSSION

The purpose of this study was to determine whether student presentation grades and feedback quality were affected by speaker placement and rater gender. Three research questions were used to test the presence of these relationships. Specifically, research question one asked whether student ratings were affected by speaker placement and proximity to presentations of various levels of quality. Findings from this study demonstrated significant differences across each of the presentations used in this analysis and the emergence of two patterns of rater error. First, ratings for *A* presentations significantly declined when preceded by lower quality presentations. Similar findings were obtained when examining the decline in ratings for *C* presentations. Second, a number of ratings for *B* and *D* presentations experienced significant increases when initiating the speaking order and when variability across presentation quality existed (e.g. *A, B, C, D, A, B, C, D*).

A variety of parallels to existing research on rater error emerge from this analysis. First, these findings support the assumption that student presentation grades are not only influenced by the quality of the presentation given by the student, but they are also influenced by the speakers' placement in a particular speaker order. Further, the quality of the presentations surrounding a particular speech significantly influenced ratings provided by undergraduate raters. This conclusion was true for both *A* and *D* presentations which experienced a significant decrease and increase respectively by raters. Results partially support the existence of both positive leniency and negative severity when variability across speakers occurred (Bock & Bock, 1981; Engelhard, 1994). In these instances the evaluators were more likely to grade high quality speeches more severely and lower quality speeches more leniently. Both sets of *A* and *C* presentations experienced significant declines in ratings when preceded by lower quality presentations. This finding suggests that raters had a difficult time making distinctions across presentations of different quality, and as a result, their final evaluations were skewed both positively and negatively. These findings also support the existence of primacy and recency effects. Raters appeared to be influenced by those presentations that appeared earlier in the speaker order. These findings have a number of parallels with previous research including Anderson and Barrios (1976) and Miller and Campbell (1959) who concluded that primacy and recency effects exist to the extent that speaker order had an impact on final grade assignment. However, this study is inconsistent with Benson and Maitlen (1975) and Turman & Barton (2003) who found

no significant relationship between rank and speaker position. When examining the mean scores for all speakers as a whole, central tendency appeared to occur across raters for each group (Saal, Downey, & Lahey, 1980). Presentation scores across the eight speakers were relatively low ranging from 78.67 (11.40) to 64.36 (12.63).

There are a number of implications for the above findings concerning rater error and speaker order. First, these findings demonstrate that evaluating eight speeches of varying quality at one time could increase the likelihood of rater error happening if a particular combination of speaker placement occurred. As a result, it seems evident that the circumstances of these various speaking situations limit the rater from making an accurate assessment of the speaker's performance. Second, these findings might suggest the need for additional assessment to take place in those performance-based classrooms where class size remains high. Peer assessment is one particular method that raters could use to assist in determining accuracy of performance assessment. Research examining the use of peer assessment as a function for analyzing student presentations has been addressed by a number of researchers with mixed results. MacAlpine (1999) and Orsmond, Merry, and Reiling (1996) obtained correlation coefficients in the ranges of .80 and .74 respectively when utilizing a likert scale assessment tool for students to complete. Kwan and Leung (1996) however found unacceptable correlation coefficients ( $r = .20$ ) when having students provide raw scores, and Freeman (1995) obtained limited success with the use of peer team/groups ( $r = .26$ ). However if appropriate training and assessment tools are util-



ized, peer assessment could assist in checking the accuracy of scores provided by raters (Bock & Bock, 1981). One avenue for future research could be the examination of similarities across peer and instructor assessments and the impact similarities/dissimilarities would have on perceived instructor credibility. Third, these findings could provide justification for a type of error referred to as “systematic distortion” (Carlson & Mulaik, 1993, p. 111). Carlson & Mulaik (1993) argue that when individuals make assessments of others they:

. . . develop common, implicit notions about “what goes with what” based on the conceptual or semantic similarities among attributes. When people are asked to make memory-based judgments of previously observed trait or behavior attributes, the ratings are systematically biased in the direction of the conceptual similarity schema....ratings of human attributes are merely linguistic artifacts that have little, if any, relation to true behavioral covariance. (p. 88)

In the context of making speech evaluations across a number of speakers the order and quality of the presentations ultimately impacts a rater’s ability to make distinctions across presentations (e.g., the first and second presentations both had good introductions and as a result they are scored alike). Thus the idea that similarities in the presentation directly preceding and following a speaker could impact the rater’s assessment is of significant importance and requires additional analysis.

No significant differences were found when examining the impact of speaker order on the quality of written feedback to students in research question two. However, one should note that the potential fatigue associ-

ated with written feedback may not be as evident after only eight presentations. Proving that fatigue is a cause of poor feedback would require a much larger and more inclusive research design than the current study could accommodate. Although this study used well-trained raters, they are still largely novice. Even with the novice label, it is unlikely that fatigue would be evident with eight speakers in one isolated speech day. Placing these same raters in the context of a typical faculty experience where two or three sections of the course are taught by the same instructor and speakers from all sections speak on the same day is much more likely to reveal evidence of fatigue. This means that a more longitudinally focused study needs to be done that tracks this issue over the course of a semester.

The third research question focused on determining whether rater gender would influence the quality of comments students received for their respective presentations. Findings indicated that females provided written comments of higher quality than male raters; however, only slight differences emerged across these two groups. The minor differences in feedback quality may have been a result of selection procedures when choosing both male and female speakers of similar quality for raters to grade. Research has suggested that raters are more likely to rate students of the same sex more highly, and by averaging the scores across the four male and female speakers may have hindered our ability to obtain large differences in feedback quality. Moreover, power was significantly reduced when including speaker sex into the analysis of rater sex differences.

Findings from these research questions do answer a number of concerns in regards to the quality of rater

feedback in the performance-based course. The assumption that rater feedback would decline as speaker order increased was disproven, indicating that quality feedback was provided across all speakers. A significant issue emerges from this and previous findings. Quigley (1998) pointed out that feedback on oral assignments benefits students most through “clear grading criteria, structured practice and specific feedback” (p. 48). However, these analyses suggest that not only were raters influenced by speaker order and quality when assigning scores, but they also appeared to be able to provide written justification for those scores. One must consider how raters justify the grades they assigned in those instances where significant increases or decreases in ratings occurred. Book (1985) found that an improvement in speaking skills is directly related to effective feedback “in accordance with the assignment” (p. 22). Future research examining the implication of speaker order and evaluation quality could attempt to determine how lower scores are justified to speakers. In situations where scores were reduced, feedback could ultimately cause a decline in presentation quality in the future.

Despite the findings obtained in this analysis, there are a number of limitations that must be considered when interpreting the results from this study. First, even though extensive training occurred to familiarize raters with appropriate assessment methods, undergraduate students were used in this analysis. There is some evidence to support the idea that less experienced evaluators may be more prone to experience rater error (Young, 1974). Second, because an incomplete experimental design was utilized for this analysis, the selection of the speaker placement for each group may cause

the findings to over represent the potential of this phenomenon. A complete experimental design would have required an additional 56 groups to achieve that total number of possible speaker combinations. From this analysis each of the groups demonstrated significant differences for at least one of the eight speeches and the percentage could drop significantly if a complete experimental design was performed. Third, the fact that raters had a difficult time making distinctions across presentations of varying quality may have been a result of the training procedures. Because raters were trained by evaluating individual presentations during each training session, rather than multiple presentations, may have had an impact on their ability to make clear distinctions across speakers. Finally, because raters were not required to interact with these speakers in the classroom, there may be some logic to suggest that they felt less inhibited in providing feedback and assigning overall scores. Watching speeches on videotape is not the same as a live experience in terms of the overall critical distance the mediated version provides. However, because raters had no previous contact with the presenters prior to assessment, the potential impact of the halo effect was eliminated as a type of rater error that may have emerged.

Despite the above limitations, this study does have a number of practical implications for the basic course director. Although undergraduate raters were utilized, the training sessions made use of many of the same training procedures employed by basic course directors when training graduate teaching assistants. The findings suggest that GTA's should be trained to understand the increased potential for rater error once fluctuations in

speaker quality exist. Furthermore, using training methods which focus on evaluations of single presentations followed by discussion may serve to increase the potential for rater error because this procedure does not accurately reflect what new GTA's will face during a typical presentation day. Finally, directors who are faced with the decision to increase the number of speeches given by students in a given class period, must consider not only the pedagogical implications, but also the potential unfair advantage it places on the effective evaluation of student presentations. This study could potentially serve as a rationale for maintaining current course structures when administrative pressure begins to emerge.

This study has demonstrated that when grade variability exists for a group of speakers, the placement of those speakers can significantly affect the final grade students are assigned. When examining previous research utilizing a similar experimental design (Turman & Barton, 2003) with only four speakers and presentations of similar quality, no significant differences were obtained. Including four additional speakers, and better reflecting a typical speech day with inconsistent presentation quality caused grade assignment across groups to change based on speaker order. Although future research needs to be done, this study does show some promise in terms of the impact increased class size could have on student learning and their right to receive fair and accurate assessment. In addition, these findings should be valuable for administrators who insist that increasing class size is the first option for reducing costs in the basic course. In the face of increasing demands for accountability, the more that educated planning de-

isions can be made the more likely students are to obtain a better, more equitable education.

## REFERENCES

- Agee, W. H., & Smith, W. L. (1974). Modifying teachers' attitudes towards speakers of divergent dialects through inservice training. *Journal of Negro Education, 43*, 82-90.
- Anderson, N. H., & Barrios, A. A. (1961). Primacy effects in personality impression formation. *Journal of Abnormal and Social Psychology, 63*, 346-350.
- Bannister, B. D., Kinicki, A. J., Denisi, A. S., & Hom, P. (1987). A new method for the statistical control of rating error in performance ratings. *Educational and Psychological Measurement, 47*, 583-596.
- Benson, J. A., & Maitlen, S. K. (1975). An investigation of the relationship between speaking position and rank assignment in forensic competition. *Argumentation and Advocacy, 11*, 183-188.
- Bishop, G. F. (1987). Experiments with the middle response alternative in survey questions. *Public Opinion Quarterly, 51*, 220-232.
- Bock, D. G. (1970). The effects of persuadability on leniency, halo, and trait errors in the use of speech rating scales. *Speech Teacher, 19*, 296-300.
- Bock, D. G., & Bock, E. H. (1977). The effects of sex on experimenter, expectancy inductions, and sex of the rater, on leniency, halo, and trait errors in speech

- rating behaviors. *Communication Education*, 26, 298-306.
- Bock, D. G. & Bock, E. H. (1981). *Theory and research into practice: Evaluating classroom speaking*. Urbana, IL: ERIC Clearinghouse on Reading and Communication Skills.
- Book, C. L. (1985). Providing feedback: The research on effective oral and written feedback strategies. *Central States Speech Journal*, 36, 14-23.
- Buerkel-Rothfuss, N. L. & Gray, P. L. (1990). Graduate teaching assistant training in speech communication and noncommunication departments: A national survey. *Communication Education*, 39, 292-307.
- Carlson, M., & Mulaik, S. A. (1993). Trait ratings from descriptions of behavior as mediated by components of meaning. *Multivariate Behavioral Research*, 28, 111-159.
- Cheatham, R. & Jordan, W. (1972). An investigation of graduate assistants in teaching the college public speaking course. *Speech Teacher*, 21, 107-114.
- Dunbar, S. B., Koretz, D., & Hoover, H. D. (1991). Quality control in development and use of performance assessments. *Applied Measurement in Education*, 4, 289-304.
- Ehrensberger, R. (1945). An experimental study of the relative effectiveness of certain forms of emphasis in public speaking. *Communication Monographs*, 12, 94-111.
- Engelhard, G. J. (1994). Examining rater errors in the assessment of written composition with a many-fac-

- eted Rasch model. *Journal of Educational Measurement*, 31, 93-112.
- Fedler, F., & Smith, R. F. (1992). Faculty members in AD/PR perceive discrimination in academia. *Public Relations Review*, 18, 79-89.
- Ford, H. T., Puckett, J. R., & Tucker, L. A. (1987). Predictors of grades assigned by graduate teaching assistants in physical education. *Psychological Reports*, 60, 735-739.
- Freebody, P., & Anderson, R. C. (1986). Serial position and rated importance in the recall of text. *Discourse Processes*, 9, 31-36.
- Freeman, M. (1995). Peer assessment of group work. *Assessment and Evaluation in Higher Education*, 20, 289-299.
- Gibson, J. W., Gruner, C. R., Hanna, M. S., Smythe, M. J., & Hayes, M. T. (1980). The basic course in speech at U.S. colleges and universities: III. *Communication Education*, 29, 1-9.
- Gibson, J. W., Hanna, M. S., & Huddleston, B. M. (1985). The basic speech course at U.S. colleges and universities: IV. *Communication Education*, 34, 281-291.
- Goulden, N. R. (1994). A program of rater training for evaluating public speeches combining accuracy and error approaches. *Basic Communication Course Annual*, 2, 143-183.
- Guilford, J. P. (1936). *Psychometric methods*. New York: McGraw-Hill.



- Harper, B. H. & Hughey, J. D. (1986). Effects of communication responsiveness upon instructor judgment grading and student cognitive learning. *Communication Education*, 35, 147-156.
- Jersild, A. T. (1929). Primacy, recency, frequency and vividness. *Journal of Experimental Psychology*, 12, 58-70.
- King, J. L. (1998). The effects of gender bias and errors in essay grading. *Educational Research Quarterly*, 22, 13-25.
- Knower, F. H. (1936). A study of the rank-order methods of evaluating performances in speech contests. *Journal of Applied Psychology*, 24, 633-644.
- Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly*, 51, 201-219.
- Kwan, K. P. & Leung, R. (1996). Tutor versus peer group assessment of student performance in a simulation training exercise. *Assessment and Evaluation in Higher Education*, 21, 205-213.
- Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, 3, 331-345.
- MacAlpine, J. M. (1999). Improving and encouraging peer assessment of student presentations. *Assessment and Evaluation in Higher Education*, 24, 15-25.

- Mason, K. J. (1976). *Intercollegiate individual events competitors' speaking order and rankings*. Unpublished manuscript, Nebraska Wesleyan University.
- McMillan, J. H. (2001). *Classroom Assessment: Principles and practice for effective instruction* (2nd Ed.). Needham Heights, MA: Allyn & Bacon.
- Miller, G. M. (1964). Agreement and the grounds for it: Persistent problems in speech rating. *Speech Teacher, 13*, 257-261.
- Miller, N., & Campbell, D. T. (1959). Recency and primacy persuasion as a function of the times of speeches and measurements. *Journal of Abnormal and Social Psychology, 59*, 1-9.
- National Center for Educational Statistics (1996). Total fall enrollment in institutions of higher education, by control, level of enrollment, and state: 1994. [Electronic version]. *The Digest of Education Statistics, 194*.
- O'Hair, H. D., Friedrich, G., Wiemann, G. W., & Wiemann, M. O. (1995). *Competent communication*. New York: St. Martin's.
- Orsmond, P., Merry, S. & Reiling, K. (1996). The importance of making criteria in the use of peer assessment. *Assessment and Evaluation in Higher Education, 21*, 239-250.
- Osgood, C. E., Suci, C. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana, IL: University of Illinois Press.

- Popham, W. J. (2002). *Classroom assessment: What teachers need to know* (3rd Ed.). Boston, MA: Allyn & Bacon.
- Quigley, B. L. (1998). Designing and grading oral communication assignments. *New Directions for Teaching and Learning*, 74, 41-49.
- Roach, K. D. (1991). Graduate teaching assistants' use of behavior alteration techniques in the university classroom. *Communication Quarterly*, 39, 178-188.
- Saal, F. E., Downey, R. G., & Lahe, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, 88, 413-428.
- Sato, K. (1990). Recency effects and distinctiveness of position/order information. *Perceptual and Motor Skills*, 71, 259-266.
- Sawyer, C. R. & Behnke, R. R. (1997). Technological approaches for improving grading efficiency and compatibility in multi-section/multi-instructor communication courses. *Journal of the Association for Communication Administration*, 3, 163-169.
- Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, 4, 25-29.
- Turman, P. D., & Barton, M. H. (2003). Stretching the academic dollar: The implications and effectiveness of instructor assistants in the Basic Course. *Basic Communication Course Annual*, 15, 144-168.
- Van der Vleuten, C. P., & Swanson, D. B. (1990). Assessment of clinical skills with standardized patients: The state of the art. *Teaching and Learning in Medicine*, 2, 58-76.

- Williams, D. E. & Roach, K. D. (1993). Graduate teaching assistant perceptions of training programs. *Communication Research Reports, 10*, 183-192.
- Williams, D. E. & Schaller, K. A. (1994). A cross-disciplinary study of GTAs' views on topics used in training programs. *The Journal of Graduate Teaching Assistant Development, 2*, 13-20.
- Young, S. (1974). Student perceptions of helpfulness in classroom speech criticism. *Speech Teacher, 23*, 222-234.
- Young, C. (1998, November). *Intercultural adaptation in the classroom: The ethics of grading and assessing students with minimal proficiency in speaking English*. Paper presented at the annual meeting of the National Communication Association, New York.