A Program of Rater Training for Evaluating Public Speeches Combining Accuracy and Error Approaches

Nancy Rost Goulden
Kansas State 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
Available at: http://ecommons.udayton.edu/bcca/vol2/iss1/13

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, mschlangen1@udayton.edu.
A Program of Rater Training for Evaluating Public Speeches Combining Accuracy and Error Approaches

Nancy Rost Goulden

IMPORTANCE OF RATER TRAINING

Educators in general, and public speaking teachers specifically, face the constant challenge of improving their methods of evaluating student achievement. In many beginning public speaking courses, a large portion of the student's final grade is based on grades assigned to individual speeches. In addition, many communication educators find themselves responsible for developing wide-scale speech testing programs to be used for placement or to establish student competency/achievement levels. Oral communication teachers have the responsibility to make these evaluation results as accurate (valid), consistent (reliable) and fair (both valid and reliable) as possible.

One means of improving speech evaluation is a carefully constructed program for training in the scoring of speeches. A deliberate planned program of rater training increases both reliability (Quellmalz, 11) and validity of scores by helping raters remain faithful to already established scoring criteria when rating speeches (Becker, 227). Charney (1984) writing about training of raters of written compositions explains how training creates such fidelity, "Training procedures are designed to 'sensitize' the readers to the agreed upon criteria"...
A Program of Rater Training for Evaluating Public Speeches

and guide them to employ those standards, rather than their own" (73).

OVERVIEW AND COMPARISON OF ACCURACY AND ERROR METHODS OF TRAINING

For evaluation of product/performance in both speech and writing, raters are usually trained by what is called the accuracy method. The accuracy method is also common in training raters for psychological counseling. In this method of rater training, the focus is on insuring understanding of the underlying concepts, understanding of the instrument and method of rating, and allowing raters to practice with sample products (Wilson and Griswold, 4). An alternative method of rater training is error training where raters are trained to be aware of and reduce common observer errors.

Wilson and Griswold (1985) set up an experimental study to compare the two training techniques (4-8). It was hypothesized that accuracy training leads to greater validity in rating, and that error training would increase reliability through the reduction of those errors, but at the same time reduce validity. The dual hypotheses were confirmed. Raters trained to identify and avoid errors did so, but the accuracy of their ratings was lower than those trained using the accuracy method. Just knowing what not to do was not sufficient to achieve both validity and reliability.

In the oral communications area, Becker (1970) recommends elements of both accuracy training (i.e., train raters to make finer discriminations) and error training (i.e., insist that raters avoid central tendency) (224). Since the two methods are not mutually exclusive and raters could benefit by improving both reliability and validity, a combined approach seems appropriate.
ACCURACY TRAINING

For accuracy training in evaluating speeches, Gundersen (1978) began with the trainer introducing and demonstrating the variables which were to be scored (402). This was followed by the introduction of taped speeches representing several quality levels. These “anchor” speeches may be analyzed and discussed in the group. Raters then practice scoring, interspersed with frequent discussion, until they reach consensus (Charney, 74).

ERROR TRAINING

Error training includes presentation to the raters description and examples of both errors which originate in the biases of the rater and come out regardless of the scale used and rater errors directly related to the rating scale such as central tendency error and logical error.

Bohn and Bohn (1985) discovered in their study that two types of rater bias error, leniency and halo errors, account for “the majority of the total error variance” (347). Leniency error refers to the tendency of the rater to scale all speakers too high or too low. The “hard graders” or “easy graders” may be consistent in their own ratings, but their scores will not reflect the true value of the performance they are assessing. Halo error may also be positive or negative, but it centers on individual speakers. Here the judges’ biases will cause them to rate a single speaker inconsistently high or low in relation to the raters’ evaluations of other speakers and in relation to the performance’s true value. In these rater situations, intrarater reliability may be high (the rater may repeatedly repeat the same errors), but interrater reliability and validity may be low.
The third type of rater error, trait error, is less common (Bohn and Bohn, 347) but does persist for some raters on some traits (Bock and Saine, 236, Bock and Munro, 371). Again raters may rate too stringently or too leniently on a specific trait of the speaker (i.e., eye contact) or a specific trait of the speech itself (i.e., organization). Geyerman and Bock found rater's evaluation of the trait "material" was affected by their attitude valence (9).

The two rater errors related to use of the scale, central tendency error and logical error may still occur even though rating scales are carefully designed to make it more difficult for the rater to clump related items. During the training, raters can be introduced to all of the above rater errors and monitored during practice so they become more aware especially of leniency, halo, trait, and central tendency errors which creep into their scoring.

PROGRAM OF RATER TRAINING

The following rater training program includes (1) general training for scoring speeches incorporating both accuracy and error training and (2) specific training for using an analytic rating scale and training for using a holistic rating scale.

The training session itself consists of lecture or oral reading of a training script by the trainer followed by practice rating of taped speeches using a rating scale or guide. The number of raters trained together in one group should be limited so that all raters have an opportunity to play an active role in the discussion of the "practice ratings". A training session requires between one and two hours depending on the number of anchor speeches from video tape which are viewed, scored and discussed.

Groups need to work with between two and four different speeches. The speeches should represent a range of quality
levels. Including both speeches which easily elicit consensus and those which provoke varied responses from the raters provides more thorough training. Speeches on a variety of topics presented by both male and female speakers and speakers whose appearance varies from each other will allow the raters to monitor themselves for halo errors. All speeches should have been prepared to fulfill the same assignment.

SCRIPT FOR TRAINING RATERS TO SCORE PUBLIC SPEECHES

Trainer reads the following script as raters follow silently.

General Introduction

"Assume that for every student speech, there is somewhere a perfect, exact grade. When we determine grades on student speeches, we are trying to get as close to the ideal score as we can. Obviously, we have no way of knowing what that absolutely valid grade is. So we approach the problem from the other direction. We try to eliminate or at least reduce those factors which pull our grade away from the perfect grade. The two general areas which contribute to 'error' are the grading procedures and the grader. In this case, 'error' does not literally mean a 'mistake' but is a term which represents the factors which make up the difference between the 'perfect' grade and the grade which is given."
I. General Training

"Today's training session will consist of three parts. We will consider common rater errors, the criteria for rating and the rating scale, and practice rating using the (either holistic or analytic) method.

Rater Error Training. As graders we each have preconceived ideas about what should be included in a speech and how a speaker should look and sound. At times our personal criteria may be so far from those of other trained, experienced raters of speeches that our ratings are unreliable and invalid. Since we can never directly observe that perfect grade or investigate just how it was determined, the best we can do is compare one rater's scores to the scores of a group of responsible raters. When investigators have made such comparisons, they have discovered that speech raters whose scores deviate extensively from the norm do so because those ratings reflect one or more personal biases about speech grading. Just as with the term 'error,' the word 'bias' here is not a pejorative term. It means the rater's personal preference differs from the majority of raters. If during the training session, you discover that some of your personal standards are causing your scoring to be inconsistent with the scores of other raters, try to put aside your biases when rating with a group. In your classroom, your individual criteria may be appropriate because of the unique experiences and expectations for that class. However, in fairness to your students, especially in a multi-section course, you may discover from the presentation today on Rater Errors some biases which you should be aware of as you grade your own students in your own classroom.

In Communication Education October, 1985, Bohn and Bohn reported that two rater errors, called the leniency error and the halo error 'accounted for the majority of the total..."
A Program of Rater Training for Evaluating Public Speeches

error variance' (348). They also identify a third common error, trait error.

Leniency error is the 'tendency of the rater to be too easy OR too hard on all speakers.' Although you may suspect (and even be proud of that characteristic of yourself as a grader) you cannot be sure if you are TOO EASY or TOO HARD until you have had an opportunity to compare your grading with other educators scoring the same speeches under the same conditions. Later in today's session as we rate taped speeches for practice, you will be able to compare your ratings with those of the others in the group.

If you are intrinsically a rater who is easily persuaded or always has great sympathy for all speakers, which is then reflected in your grading, you may need to guard against scoring too leniently. Some inexperienced raters who are unsure of their own ratings or are apprehensive about defending their grading, especially in a face-to-face confrontation with their students, grade too leniently to avoid problems.

Hard graders may have developed very stringent standards in the hopes of spurring their students to excellence. In their zeal, these graders may have set almost impossible levels which few, if any, students can reach. Again, today when you compare your scores to other raters, you will get an indication of whether you need to readjust your degree of 'leniency' either positively or negatively.

Halo error is the 'tendency of the rater to be too easy or too hard on a specific speaker.' Bock and Bock reported that one manifestation of the halo error is related to gender (6). Both men and women tend to give higher scores to speakers of their own gender than to speakers of the other sex. Another common halo error is inflating the grades of speakers the rater knows and likes. As you rate taped speeches of students you have never met, you may find certain physical attributes or behaviors appeal to you or repel you to the extent that those characteristics cloud your evaluation of the speech.
The rater should try to apply the same standards to all speeches and the performance of those speeches and resist the urge to compensate the speaker for whom the rater feels sorry or elevate a grade just because a speaker 'really seems to be trying.' Responding to the individuality of a speaker is part of the evaluation process. Therefore, guarding against the halo error does not mean that you must reduce grading to a sterile mechanical process that eliminates the humanity (and subsequent uniqueness) of a speaker. Again we are trying to be aware of biases that causes our rating to move out of the mainstream or be unfair to individual speakers.

Trait error is the 'tendency of the rater to be too easy or too hard on a specific trait (category) included in the scale regardless of speakers.' The traits or characteristics of speeches and speakers which are on the rating scale used for this training sessions were chosen by canvassing speech educators and surveying nine popular college-level public speaking texts. Therefore, the traits which are important to you as a rater are probably on the rating forms. however, because of your individual training and experience, you may have a list of 'have-to items,' traits that 'have to' be present in a certain form or at a certain level of excellence. For example, a rater might have decided that it is absolutely essential that all speakers orally identify main points in a preview with numbers ('first, second'). If a student deviates from the form, the speech might as well be over. It really does not matter what else happens. The rater will ignore all other items or manipulate the scoring of those items so that all other aspects of the speech or presentation have no effect on the outcome. Other raters become so concerned about specific delivery behaviors, that those items take on an exaggerated importance and block from the raters' consciousness all other traits.

The speakers you will rate during this training should be judged on the basis of their performances and speeches in relation to their assignment. You will have read what their
A Program of Rater Training for Evaluating Public Speeches

instructor emphasized when assigning the speech. If you had designed the assignment, you might have included other criteria; however for this training session, try to correlate your expectations with those of the particular assignment.

Even when raters are able to treat each trait as a discrete step in the rating process and faithfully utilize assignment criteria, they may still commit a trait error by creating idiosyncratically high standards for some traits. The raters may consider other traits so unimportant that they become toss-away items – just be sensitive to unwarranted stringency or leniency for individual traits.

Raters may also make trait errors when they let their rating of one trait carry over to other traits because those characteristics are located near each other on the rating scale. In a similar fashion, raters may unintentionally group characteristics which are related to each other (for example, all items which deal with delivery or content) and rate them all the same. (For analytical training: We have tried to organize traits in the order that we assume you will observe that item during the presentation of the speech, rather than in groups of logically related traits.) (For holistic training: This type of trait error is less a factor in holistic rating since you will not be rating individual characteristic separately.)

As we practice rating, try to be aware of your tendencies toward these rating errors. Do not become so overly concerned that you are immobilized as a rater or begin to see faults where none exist. Because of the opportunity to compare your rating with the other raters in your group, you and the trainer may be able to identify the presence of errors which can be reduced and will make you an even more reliable and valid rater. On the other hand, you may find that your rating is relatively free of biases."

Volume 2, November 1990
<table>
<thead>
<tr>
<th>Persuasive Speech Characteristic</th>
<th>Present at Excellent Level</th>
<th>Present at Good Level</th>
<th>Present at Satisfactory Level</th>
<th>Present at Inadequate Level</th>
<th>Absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Thesis statement tell what the audience should believe and/or do.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Speaker shows problem is widespread and/or severe.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Speaker defines necessary terms.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Speaker related problem to the audience.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Speaker employs natural delivery style (e.g. conversational, interactional).</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arguments are sufficiently supported with evidence (e.g. relevant, credible, recent.)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speaker shows advantages of solution.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speaker maintains eye contact.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speaker cites sources.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speaker projects confidence.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speaker states costs and/or benefits for the audience.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speaker uses language choices that accomplish the speaker's purpose.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speaker refrains from distracting delivery behaviors.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speaker addresses potential audience objections.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
15. Speaker avoids weak arguments that do not contribute to acceptance of thesis statement.

16. Speaker's treatment of issue is responsible (e.g. honest presentation, concern for audience safety and welfare).

17. Conclusion reinforces thesis statement by means of summary and/or appeal.

18. Speaker's verbal and nonverbal messages reveal commitment to proposal.

19. Structure of speech is clear (e.g. preview, transitions, summaries).

20. Speaker's main points are organized in a persuasive pattern.
21. Speaker uses voice and body to increase understanding and/or emotional impact.
22. Speech fits requirements of assignment (e.g. conviction, actuation, length).

<table>
<thead>
<tr>
<th>Level</th>
<th>Number of Items Marked</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td>x 4</td>
<td></td>
</tr>
<tr>
<td>Good</td>
<td>x 3</td>
<td></td>
</tr>
<tr>
<td>Satisfactory</td>
<td>x 2</td>
<td></td>
</tr>
<tr>
<td>Inadequate</td>
<td>x 1</td>
<td></td>
</tr>
<tr>
<td>Absent</td>
<td>x 0</td>
<td>0</td>
</tr>
</tbody>
</table>

Total 22

Figure 1. Analytic Rating Scale
II. Analytic Training

"Analytic scoring of a speech means that the rater records a core for separate characteristics or traits of a speech and the performance of that speech. These individual scores are then combined to produce the overall grade for the speech. The underlying assumption of analytic rating is that a speech is the sum of its parts and by carefully separating the speech into those component parts for consideration during scoring and then recombining the scores, the rater will produce a fair, accurate grade which summarizes the speech and performance.

Accuracy Training. Now please look at the analytic rating scale. There are twenty-two statements describing the elements which make up a classroom persuasive speech and then presentation of that speech. For each speech trait, decide, based on the taped presentation, if the speaker has included that characteristic. You will then determine the degree of the speaker’s expertise for the traits present.

At this time, please read through the twenty-two items to make sure that you understand the terminology. Do not yet concern yourself with the levels. (Pause).

Look at the first item. Please note that a ‘thesis statement’ does not have to be restricted to a rigid word formula but is the speaker’s overt statement of the central idea, proposition or claim.

Now read items two and four. The term ‘problem’ does not have to be used by the speaker orally. A ‘problem’ represents any situation which the speaker advocates to be changed. And ‘solution’ (item seven) is the change the speaker advocates.

Next look at item eleven. ‘Costs’ may include disadvantages or harms which impact on the listeners either as a result of the problem or the solution. ‘Benefits’ are usually advantages resulting from the proposed solution.
Now read item twenty. Traditional ‘persuasive patterns’ include: problem-solution, cause-effect, statement of reasons, comparative advantages, criteria-satisfaction and motivated sequence.

Now look through the list of items again and consider how to determine if the trait is present. (Pause).

To distinguish the difference between the presence of a trait at the Excellent and Good Levels, assume that the Good Level means that the speaker has met expectations. The Excellent Level should be awarded when the speaker has gone above and beyond expectations. The Good Level is the ‘touchstone.’ Once you have established that standard in your own mind, you should be able to move down to the Satisfactory and Inadequate Levels and up to the Excellent Level.

At this time, go through the list one more time and visualize each trait at the Good Level. (Pause).

Do the same for the Inadequate Level. (Pause).

We are now ready to begin evaluating a speech. The taped speeches are actual classroom speeches from undergraduate university classes. Please review the description of the speaking assignment. (Pause).

You may mark the rating scale either as you listen to the speech or when the speech is finished. Make a check-mark for the level you have chosen for each trait. Count the number of marks for each level and multiply by the factor given on the score sheet on the last page of the rating scale. Then total those products. Record your total score. The purpose of producing a total is so we can compare scores in order to help you determine if you need to adjust your standards of scoring.

Scores from the first speech are collected, and the mean and range calculated and announced. Each rater then reports the numbers of the items from the rating instrument which that rater marked at the highest level. In discussion, raters defend their choices. Opportunity is provided for clarification of specific items. Then all raters report the items which they
scored at the lowest level. These choices are also discussed. The same process is repeated for a second speech. Raters are reminded to be aware of their own rater errors and try to adjust their expectations to correspond with those of the group.

The group may need further practice with a third or fourth speech.

III. Holistic Training

"Holistic rating means that the speech is considered as a whole and that the rater assigns only one score to represent the content of the speech and its presentation. The rater does not record any subscores or mark specific characteristics of the speech or speaker. However, the score is not just an arbitrary number drawn out of the air but is the result of matching the speech the rater has listened to with written descriptions of speeches at various levels. However, since the descriptions are rather brief, the rater may automatically factor in characteristics which are not included in the descriptions. Holistic rating is based on the assumption that a speech is more than the sum of its parts and that no rating scale listing the component parts of a speech is complete. Holistic rating therefore can accommodate the unexpected and also allow the rater the latitude to reward uniqueness within the framework of general criteria.

Accuracy Training. Look at the descriptions of speeches at the four levels from Excellent to Inadequate. As you see, the descriptions have been divided into five categories to help with the comparisons. First read the total description of the Excellent Level. (Pause)
### EXCELLENT LEVEL
Between 80 and 61 points

**AUDIENCE AND SITUATION ADAPTATIONS:**
Speaker precisely meets requirements of assignment; explicitly points on legitimate relevance and application of issues, problems, solution, and advantages to audience; adopts a responsible position relative to audience.

**ORGANIZATION:**
Speaker chooses and orally presents, in such a way that audience can effortlessly follow, a pattern of organization that enhances the persuasive purpose of the speech and completely supports the thesis statement.

**CONTENT:**
Speech leads to inevitable acceptance of speaker’s proposal by speaker’s presentation of strong, relevant, compelling, valid arguments; an abundance of credible evidence, and powerful emotional and psychological appeals.

**LANGUAGE:**
Speaker makes language choices which are unfailingly clear, precise, accurate and increase interest and emotional impact of message.

**DELIVERY:**
Speaker uses extemporaneous, natural delivery style; projects confidence and sincerity. Voice, body movements and eye contact result in direct and effective connection with audience; speaker avoids distracting behaviors.

### GOOD LEVEL
Between 60 and 41 points

**AUDIENCE AND SITUATION ADAPTATIONS:**
Speech is within time limits and matches the characteristics of the type of speech assigned. May not be stated explicitly but for the most part, the problem solution, and advantages are relevant and applicable to the audience. Speaker adopts a responsible position relative to the audience.
ORGANIZATION:
Speaker aids the audience in following the structure of the speech; pattern of main points is consistent with persuasive purpose and for the most part supports the thesis statement.

CONTENT:
Speech opens the door for acceptance of speaker’s proposal because of speaker’s presentation of valid arguments, credible evidence and psychological appeals.

LANGUAGE:
Speaker’s language choices contribute to clear understanding; occasionally unique choices increase interest or emotional impact of message.

DELIVERY:
On rare occasions, speaker switches from extemporaneous mode to manuscript or memorized mode. For the most part, delivery is natural and speaker only infrequently shows a lack of confidence or control; voice behaviors and body movements are not major distractions but fail to enhance presentation.

SATISFACTORY LEVEL
Between 40 and 21 points

AUDIENCE AND SITUATION ADAPTATIONS:
Assigned speech requirements and characteristics do not match perfectly with this speech. Some, but not all, aspects of the proposal are directly applicable to this audience. While not blatantly irresponsible, speaker’s position does not obviously have audience safety and welfare as primary considerations.

ORGANIZATION:
Speaker adheres to a planned structure for the speech which audience can follow although the speaker has not presented obvious previews, summaries or transitions. Some main points may be tangential to the persuasive purpose and the development of the thesis.
CONTENT:
Some arguments and appeals of the speech are acceptable to the audience. Most arguments are valid and supported with some evidence; speaker has included few acceptable emotional or psychological appeals.

LANGUAGE:
Language choices are utilitarian. Occasionally speaker diminishes impact by vagueness or inappropriate choices.

DELIVERY:
Speaker consults notecard frequently; there is some evidence of programmed or stylized delivery or a lack of confidence or control. Occasional vocal problems such as vocalizers, lack of fluency, lack of crisp articulation may be present. Speaker engages in purposeless body movements or remains stiffly rooted in one place.

INADEQUATE LEVEL
Between 20 and 0 points
Speech obviously violates assignment constraints; topic or treatment are not appropriate since audience is already in full agreement with speaker's stance or the problem, solution, advantages do not apply to this audience; speaker asks audience to take action which poses threat to audience safety or welfare.

ORGANIZATION:
Speech does not appear to be organized into cohesive discrete blocks under main points. The ideas and their sequence are inconsistent with the speaker's purpose and do not develop the thesis.

CONTENT:
Arguments are weak or fallacious; evidence is lacking or based solely on speaker's personal opinion; emotional and psychological appeals are either absent or dominate to the exclusion of rational appeals.
LANGUAGE:
Speaker uses vague, general language; cannot pronounce or obviously does not understand some terms; uses gender or ethnic terms which show lack of sensitivity to audience; depends heavily on jargon.

DELIVERY:
Speaker reads or recites speech from memory; vocal delivery patterns (repetitive rhythms, pitch, rate, volume) make it difficult for audience to understand or listen to speech; speaker makes little direct contact with audience; distracting behaviors pull listeners away from message.

Figure 2. Holistic Rating Scale

Now read the category of Audience and Situation Adaptations for all levels. (Pause).

Now read the category of Organization for all levels. (Pause).

Now read the category of Content for all levels. (Pause)

Now read the category of Language for all levels. (Pause).

Now read the category of Delivery for all levels. (Pause).

Keep in mind that although we have just looked at the component parts of speeches, you will score the speech as a whole. Do not record any subscores. You may hear and see speeches which are strong in some categories but weak in others. Your job is to find the overall description of the speech which best matches the speech you see and hear. Notice that a range of scores is given for each level. If the speech you are rating matches the description perfectly in all aspects, you would choose a score at the top of the range. If the speech fits the level in a general way, but the fit is not perfect, then you will choose a score within the range but not at the top.

One procedure for scoring holistically is immediately after the speech decide on the general level; then reread that level to verify the match. If the speech matches most categories but
obviously not all, do not automatically assume you must move the speech to a lower category. The question is: as a whole, where does the speech fit?

We will not practice rating a taped speech. The speeches we will view and score are classroom speeches taped during presentation for the students' class. Now look at the description of the assignment. (Pause).

As you listen to the speech, look at the descriptions and make tentative judgments, but do not write any scores down. As soon as the speech ends, quickly decide on and record your score while the overall impression is fresh.”

Scores from the first speech are collected, and mean and range calculated and announced. Individual raters are asked to defend the level of the score they selected by pointing to descriptors of the speech at that level from the holistic instrument. Raters have the opportunity to disagree and support their viewpoint or ask the trainer for clarification. The process is repeated for another speech. Raters should be reminded of rater errors and the need to try to adjust their expectations to correspond with those of the group.

If the trainer determines that the group needs more practice or has not moved toward consensus, a third and/or fourth speech should be viewed and discussed.

The training program can be adjusted for different rating instruments or methods of scoring speeches. Raters can be trained in either the analytic method or the holistic method or both. It is better to train raters to use only one method at a time to avoid confusion. The general training used alone will slight the accuracy approach which is emphasized during the instruction for the practice rating utilizing a specific method, either analytic or holistic. If raters are given free choice as to their scores, but are using uniform criteria, segments of both analytic and holistic rating sections may be used.
RELIABILITY AND VALIDITY OF TRAINED RATERS

This program was used in Spring 1989 to train 15 raters, all graduate teaching assistants in a basic public speaking course. Each rater was trained to use both the analytic and holistic rating scales over a period of two months. The raters then independently scored nine speeches using the analytic scale and nine speeches using the holistic scale.

Raw scores for each rater were compared, using Pearson product moment correlation, to all other raters who trained at the same time to determine interrater reliability. Scores for each rater were also correlated with scores on the same speeches determined by a panel of expert judges producing Pearson $r$ to establish concurrent validity.

The mean for interrater reliability coefficients for all experimental raters was .861 with a standard deviation of .128. The mean for concurrent validity coefficients for the raters was .826 with a standard deviation of .138.

The combination of accuracy and error rater training helps speech educators to produce acceptable levels of consistency and accuracy, resulting in more representative scores for speeches for their students.

REFERENCES

A Program of Rater Training for Evaluating Public Speeches


