2011

A Grade-Norming Exercise to Increase Consistency and Perceived Consistency in Grading among Public Speaking Instructors

Bessie Lee Lawton
West Chester University of Pennsylvania

Mary Braz
West Chester University of Pennsylvania

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/vol23/iss1/7

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.
Many colleges and universities offer a basic communication course for undergraduate students. These courses could be a hybrid public speaking and interpersonal/mass/organizational class, or they could be general education courses that have a public speaking orientation (Pensoneau-Conway, Maguire, & Paal, 2007). Engleberg, Emanuel, Van Horn, and Bodary (2008) report that when they surveyed 290 community colleges, 82.1% of respondents who had general education communication courses said the course had a public speaking focus. As departments increase the number of sections for the same course, they have had to hire adjuncts and graduate teaching assistants to supplement the regular faculty (Turman & Barton, 2003, Sawyer & Behnke, 1997). One important question to ask, therefore, is whether these instructors have acceptable levels of similarity in course content and grading.

Institutions have resorted to the standardization of courses to try and make sure the learning experience is
the same or close to the same for its students. In the effort to standardize, communication departments have used one or more of the following for general education courses: a common textbook, a common syllabus, common speech requirements, and common evaluation forms. Researchers have criticized the trend toward standardization, explaining that it takes away teacher autonomy and assumes that the same educational experience can be had by diverse students (Morreale, et al., 2006; Zompetti, 2006, Pensoneau-Conway, et al., 2007). Nevertheless, Pensoneau-Conway et al. (2007) argue that the trend is toward standardization because institutions have to justify the budget allotted to these general education requirements, which means the courses regularly go through some form of assessment. One of the components of assessment involves tracking student grades as a measure of student learning (Pensoneau-Conway et al., 2007). Often, course grades include an objective component (exams) and a subjective component (speech performance).

The issue of consistency in grading among teachers with diverse experience levels and backgrounds is problematic. For example, Engleberg, Emanuel, Van Horn, and Bodary (2008) mention that some faculty members teaching communication do not have degrees in communication. Instead, they hold degrees in English (61%) or theatre (53%), and have limited background in teaching basic communication courses. They also state that 76% of responding colleges had more part-time than full-time faculty. Anderson and Jensen (2002) report that inexperienced raters tend to give higher grades regardless of speech level (A speech or C speech). Thus, varying levels of experience and backgrounds raise the
question of whether faculty are grading in a consistent manner.

**Grading Consistency**

For decades, researchers have raised the issue of consistency in grading subjective performances such as a speech (Clevenger, 1962; Bostrom, 1968; Applbaum, Carroll, Robbins, & Stein, 1972; Littlefield, 1975; Goulden, 1990; Carlson & Smith-Howell, 1995; Behnke & Sawyer, 1998; Mottet & Beebe, 2006).

McNamara and Bailey (2006) describe how speech language pathology programs have developed portfolio-based assessments because the traditional assessment procedure of using direct observation to evaluate student performance assumed an unprejudiced judge, which quite often does not turn out to be the case. Turman and Barton (2004) mention three factors that could affect subjective judgments, namely: scoring procedures, assessment tools used, and rater bias. Of these three, the biggest source of error is rater bias. For example, Wade (1978) and Rubin (1990) found that teachers’ ratings of subjective work could be affected by student names, race, gender, handwriting, and the instructor’s perceived attractiveness of the student. Miller (1964) found that the instructor’s previous training and his/her attitude toward the topic affected how he/she rated a speech. Anderson and Jensen (2002) found that experience could even affect how instructors interpret evaluation forms.

Another issue that affects grading is whether raters use norm-referenced or criterion-referenced guidelines. Norm-referenced grading involves comparing the stu-
dent with a given population. Behnke and Sawyer (1998) mention that students could be compared to national norms; however, a more likely reference group might be students who have taken the class in the last three years. This raises the question of what a new instructor would use as a reference in the absence of teaching experience. Another form of norming might involve “curving” grades in a particular class. Thus, instructors could be told to give mostly B’s and C’s, and to reserve A’s and D/F’s for a small percentage of students. Standards would then vary from class to class even within the same institution.

Criterion-referenced grading involves grading a student on whether he has achieved a certain performance standard. In an effort to standardize grading, many departments use a common evaluation form that often contains content, organization, physical delivery, and vocal delivery components (Carlson & Smith-Howell, 1995). Anderson and Jensen (2002) concluded in their study of evaluation instruments and rater experience that evaluation forms that clearly specify criteria and have clear instructions are critical, especially for C speeches. Turman and Barton (2003) also emphasize that criterion-based grading is essential to reduce rater differences in grading. And Meyer, Kurtz, Hines, Simonds, and Hunt (2010), in their study on assessing preemptive arguments, state that having specific guidelines for instructors on how to use and interpret rubrics can help increase rater reliability.

Goulden (1992) discusses four classifications of how raters could assess. Criterion-referenced grading in common speech evaluation forms, in practice, would fall under Goulden’s first two models—atomistic and ana-
lytic assessment. These are considered more objective and therefore less subject (though not immune) to rater bias (Mottet & Beebe, 2006). Atomistic assessment looks at the presence or absence of a behavior. Analytic assessment does not just quantify presence or absence of the behavior, but rather, judges the quality of the behavior being evaluated. The question remains, however, as to what standard judges are using to evaluate the quality of specific behaviors. Most likely, they are drawing on their own experience with students, either overall or from the specific institution they are in. In this sense, the issue of norm-referenced judgment becomes relevant even in so-called criterion-referenced grading using analytic assessment.

The last two assessment models—holistic and general impression—are more normative, subjective, and therefore highly prone to bias (Mottet & Beebe, 2006). The holistic assessment model considers the performance components without grading them, and then comes up with a judgment on the overall quality of the work. Finally, general impression evaluations, the most subjective of these models, are not guided by common criteria but by the personal criteria of the rater.

In this mid-Atlantic university, instructors generally use the analytic assessment model. They evaluate student performance on each of the components on the common evaluation form as “excellent,” “competent,” or “needs improvement.” Components include organization (attention-getter, thesis, preview, main points, transitions, summary, clincher), content (adapts to audience, variety of supporting materials, source citations, language choice, and presentational aid), and delivery (appearance, eye contact, facial expression, gestures, notes, notes,
Instructor Grade Norming

stance/movement, rate, volume, enunciation, conversational tone, confidence, and enthusiasm). They also give an overall grade, but the determination of this overall grade is highly subjective. What happens is that a student is judged on a variety of components, and his final score could be either a summation of individual component scores, or a more analytic judgment as to how many of these he has achieved at acceptable levels.

It could be argued that grading, even using clearly identified criteria, has some element of norming, because faculty members have to draw on their own judgments of student performance. Each instructor’s norms are often different from the others, based on factors such as length of teaching experience and the variety of institutions he or she has taught at. Institutions have considered different methods to reduce inconsistencies in grading practices. Sawyer and Behnke (1997) describe how computer document-modeling software has been successfully used to improve the quality of instructor feedback while reducing the time needed to generate it. Behnke and Sawyer (1997) state that regular meetings between instructors and basic course supervisors are often necessary to increase comparability among instructors. Rubin (1990) also underscores the importance of rater training. Carlson and Smith-Howell (1995) found that speeches could be “evaluated reliably and validly using different evaluation forms as long as the forms address the age-old constructs of content and delivery, (but) novices tend to grade more harshly and inconsistently than experienced evaluators at first” (1995, pp. 93-94). This implies that some form of training or “bringing up to speed” is necessary to increase grading
Instructor Grade Norming

comparability among instructors with different experience levels.

Stitt, Simonds, and Hunt (2003) used a one-group pretest-posttest design to assess whether training improved inter-rater reliability among new graduate assistants using a criterion-based assessment rubric. Results showed increased reliability in scores after training. Institutions therefore train instructors to try and bring norms closer and reduce grade differences, but often, this is a one-time training event.

Therefore, questions remain as to whether regular training provides additional benefits with regard to reducing grade dispersions, and whether instructors who have several years’ experience teaching will benefit from continual training. Specifically, does it make sense to hold more than one training session at the beginning of the semester or year? At the National Communication Association meeting in 2009, the authors asked the audience how many training sessions were given to adjunct faculty or graduate teaching assistants teaching a public speaking course, and range of answers was from 0-1. Is there any benefit in terms of grading consistency if people receive more than one training experience? This study seeks to help answer this question.

**SELF-EFFICACY AND PERCEIVED NORMATIVE BEHAVIOR**

This study also looked at self-efficacy and perceived normative behavior because these are related to instructor performance. In the institution where this study was conducted, adjunct and new faculty members often talk
in meetings about how they were not sure whether their grading behaviors were in line with others. In other words, there was uncertainty regarding their own abilities (self-efficacy) and whether or not they were grading consistently with other instructors (perceived normative behavior). In addition, the institution was actively assessing the general education courses, and an important component of the assessment involved comparing student grades. If instructors are not grading in a fairly consistent manner, then comparisons across classes cannot be done.

Yilmaz (2009) explains that self-efficacy affects teacher performance in several ways. Teachers with high self-efficacy believe they can teach effectively, do their job willingly and affectionately, believe they can establish communication with and teach problematic students, and have high expectations for student success (p. 506). Young and Bippus (2008) mention that self-efficacy perceptions are related to anxiety and may influence subsequent behaviors toward the tasks required in one’s job. For example, it is related to perseverance, adaptability, and the degree of effort to teach more effectively. Thus, higher self-confidence is related to a desire to do one’s job even better. They state that graduate teaching assistants usually do not have high levels of confidence that they can do their jobs effectively, and present results of a three-day training program that increased “prosocial behavioral alternation techniques (p. 116).”

Likewise, perceptions of behavioral norms also influence people’s behavior. Sherif (1936) discussed norms as mutually negotiated rules that govern social behavior. These rules are shared belief systems surrounding a
particular behavior. As instructors develop perceptions about how other instructors are grading, they develop perceptions about behavioral norms surrounding grading. Outside grade-norming sessions, instructors are rarely given opportunity to form perceptions of how others are grading and whether or not they are grading consistently with others. Because norms are mutually agreed upon, communication is critical for perceptions of grading norms to form and to influence behavior (Latané, 1996). Instructors who are given the opportunity to discuss how and why they rendered certain grades should be more likely to develop perceptions of normative behavior surrounding grading. Thus, instructors who receive training should be more likely to feel as if they are grading consistently with others.

Given the importance of the issue of grading consistency in institutions that offer public speaking as a basic communication course, a grade-norming study was conducted in a mid-size Mid-Atlantic university that aimed to train public speaking instructors on speech grading and on the use of a common speech evaluation form. In this mid-Atlantic university, instructor training usually involved a general meeting before each semester to go over course policies, but there was no follow-up throughout the semester, and no conscious effort to provide continuous training to faculty. New instructors were therefore at a disadvantage and were left to learn as they went along. Many adjunct instructors in this university also taught at other institutions, and they have shared that there is no regular training provided in the other places, either. Thus, the issue of whether regular, continual training is needed remains critical.
today with the institutionalized importance of assessment and accountability.

Another goal of the study was to explore whether the exercise improved instructors’ self-efficacy and group normative behavioral perceptions. Since the study is about grade norming, it makes sense to look at instructors’ perceptions of how they graded compared to others and their confidence in their ability to grade. It also makes sense to look at whether actual grades were related to normative behavioral perceptions. In other words, if instructors perceived that they were grading in a consistent manner with others, to what extent did the actual grades given reflect this perception? Therefore, instructors also answered an instrument on perceived self-efficacy and normative behavioral perception questions.

While training involved discussion of the evaluation form as well as how final grades are determined, this study focuses on grades given by instructors, not on how they rated specific components of the evaluation form. Because the goal of the training was to improve consistency in grading, instructors’ perceptions that they are grading in the same way, and self-efficacy perceptions, the study had the following hypotheses:

H1: Variance among scores given by members of the grade-norming group will decrease over time compared with scores given by members of the control group.

H2: Instructors in the training group will report increased perceived agreement over time for the way in which they grade compared with instructors who do not receive training.
**Instructor Grade Norming**

H3: Normative behavioral perception of agreement with others will be positively associated with perceived self-efficacy of grading at each time period.

**METHOD**

**Design**

This study employed a 2 (training and non-training group) x 4 (four time periods) mixed groups design, with time as the within group measure and training group as the between groups measure. All data was collected anonymously in order to help minimize demand characteristics.

**Participants**

Fourteen public speaking instructors at a midsized university in the Mid-Atlantic served as participants in this study. Three of the instructors were men and 11 were women. Instructors had taught the public speaking course for an average of 11.07 semesters total (sd = 7.59) and for an average of 5.93 semesters at this particular university (sd = 3.20).

**Procedure**

The investigators explained the study during the beginning of the academic year orientation of Public Speaking instructors. The investigators then asked each one if they would be willing to take part. Everyone...
agreed, but self-selected into either experimental or control group depending on availability to attend training sessions, a caveat for interpreting the results. Because of this self-selection, it was important to check for comparability between groups. Independent samples t-tests were run to see whether there were differences in the mean scores given by each group for each time period. Results are presented in the measurement section that follows. The goal was to show that the two groups were comparable overall in how they graded. Results confirmed that there were no significant differences between training and control group in overall grading in any of the time periods. This starting point allowed us to focus on assessing grading dispersion between groups. There were seven participants in each of the training and the control group.

Treatment Group

Instructors in this group met four times over the course of a semester. Each meeting started out by having instructors evaluate two speeches independently using the departmental evaluation form, and then they filled out a questionnaire on perceived efficacy, group comparisons, and normative behavioral perceptions. The common evaluation form contained organization, content, and delivery components. Thereafter, a discussion ensued whereby members explained how they judged elements of the speech on the evaluation form, and why they gave the final grade they assigned. They discussed why each speech deserved the grade they gave. Therefore, they came to some agreement on what constituted an A versus a B, C, or D speech. The discussions lasted
from one to one-and-a-half hours. The goal at the end of the project was to have greater understanding of all elements of the evaluation form, as well as why other instructors gave the grade they did.

Participants were asked to evaluate videotaped speeches given by students in Public Speaking courses. Students signed a release form before giving these speeches. Four speeches were informative, and four were persuasive speeches. Instructors were asked to submit speeches, which were placed into a centralized library on the department website. The study investigators selected two speeches from four instructors’ sections for the study.

**Control Group**

Instructors in this group also evaluated the same two speeches independently on each of four time periods during the semester using the departmental evaluation form, and then they filled out the same questionnaire on perceived efficacy and normative behavioral perceptions. They did these four rounds at the same period that treatment group members were having their meetings. They did not have the benefit of any meetings or discussions with other instructors.

**Measurement**

*Graded Scores.* Participants graded two speeches in each of four time frames for a total of eight speeches. Grades were measured on a continuum ranging from 0 to 100. Graded speech scores were standardized in order to allow us to conduct analyses on dispersion rates. This
step was necessary because we are not testing for differences in speech quality (some speeches might have been better than others, e.g. an “A” quality speech versus a “C” quality speech, and therefore deserving of a higher grade than others) but rather we were testing the degree of dispersion of scores around the mean score. Means and standard deviations for graded speech scores are reported by group in Table 1.

### Table 1
Means and Standard Deviations for Speech Grade by Group

<table>
<thead>
<tr>
<th></th>
<th>Training Group</th>
<th>Non-Training Group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Speech 1</td>
<td>80.29</td>
<td>5.41</td>
</tr>
<tr>
<td>Speech 2</td>
<td>75.29</td>
<td>3.73</td>
</tr>
<tr>
<td>Speech 3</td>
<td>64.29</td>
<td>4.68</td>
</tr>
<tr>
<td>Speech 4</td>
<td>78.14</td>
<td>3.39</td>
</tr>
<tr>
<td>Speech 5</td>
<td>80.57</td>
<td>4.20</td>
</tr>
<tr>
<td>Speech 6</td>
<td>66.14</td>
<td>5.90</td>
</tr>
<tr>
<td>Speech 7</td>
<td>83.14</td>
<td>1.95</td>
</tr>
<tr>
<td>Speech 8</td>
<td>76.14</td>
<td>2.11</td>
</tr>
</tbody>
</table>

Principal Components Analysis was conducted for each pair of grades at each time period to assess communalities. Factor loadings for each of the four time periods were >.72. Thus, standardized speech scores at each time period were summed and averaged to form an index of grades at each time period.

As noted previously, we checked whether there were differences in the mean scores given by each group for
each time period. This is important to show that the two groups were comparable overall in how they graded. Results confirmed that there were no significant differences between training and control group in overall grading. The purpose of the paper instead is to see whether trained instructors were grading more consistently over time by looking at the dispersion rates around mean scores.

The results of the t-tests indicated there were no differences between training and control group grade scores for any of the four time periods. Means, standard deviations, and test statistics are reported in Table 2.

| Table 2 | Grade scores for training and control group in four time periods |
|-----------------|-----------------|-----------------|-----------------|-----------------|
|                | Training Group  | Non-Training Group | T(12) p         |
|                | Mean            | SD               | Mean            | SD               |
| Time 1         | 77.79           | 3.22             | 80.14           | 3.41             | –1.33 ns        |
| Time 2         | 71.21           | 2.38             | 69.36           | 5.96             | 0.77 ns         |
| Time 3         | 73.36           | 3.21             | 77.93           | 7.49             | –1.49 ns        |
| Time 4         | 79.64           | 1.70             | 79.43           | 3.45             | 0.15 ns         |

Perceived Normative Behavior. We developed four items to measure the extent to which instructors believed they were grading consistently with other instructors at the university. Items included, “I gave the same grade to the speeches viewed today as the other instructors did,” “Other instructors’ comments about the speeches we viewed were very similar to my own,”
“If we compared completed evaluation sheets for the speeches we viewed, mine would look just like everyone else’s,” and “Everyone here gave the same grade to the speeches that I gave.” These items were measured at each time period on a seven point scale, with higher numbers indicating greater perceived normative behavior. Confirmatory factor analysis was employed to assess internal consistency, and yielded RMSE < .10 for the four items at each time period. Thus, the indicators were summed and averaged to form a Perceived Normative Behavior Index for each time period. Means, standard deviations, and scale reliabilities for each index are reported in Table 3.

<table>
<thead>
<tr>
<th>Time</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Cronbach’s Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time 1</td>
<td>4.43</td>
<td>1.19</td>
<td>.94</td>
</tr>
<tr>
<td>Time 2</td>
<td>5.18</td>
<td>.79</td>
<td>.85</td>
</tr>
<tr>
<td>Time 3</td>
<td>5.25</td>
<td>.84</td>
<td>.87</td>
</tr>
<tr>
<td>Time 4</td>
<td>5.48</td>
<td>.73</td>
<td>.91</td>
</tr>
</tbody>
</table>

**Self-efficacy.** In order to test the extent to which instructors exhibited increased self-efficacy over time, we developed eleven items to measure the extent to which instructors believed they could grade a speech fairly. Items included, “I’m sure I can do an excellent job evaluating student speeches,” “I feel confident that I can
fairly judge all items on the evaluation form,” “I feel confident that I can judge if a student cites sources properly,” and “I can tell if a speech is organized well.” These items were measured at each time period on a seven point scale with higher numbers indicating greater self-efficacy. Confirmatory factor analysis was employed to assess internal consistency, and yielded RMSE <.10 for the eleven items at each time period. Thus, the indicators were summed and averaged to form a Self-efficacy Index for each time period. Means, standard deviations, and scale reliabilities for each index are reported in Table 4.

<table>
<thead>
<tr>
<th>Time</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Cronbach's Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time 1</td>
<td>6.05</td>
<td>.80</td>
<td>.94</td>
</tr>
<tr>
<td>Time 2</td>
<td>6.31</td>
<td>.62</td>
<td>.92</td>
</tr>
<tr>
<td>Time 3</td>
<td>6.50</td>
<td>.48</td>
<td>.92</td>
</tr>
<tr>
<td>Time 4</td>
<td>6.55</td>
<td>.38</td>
<td>.88</td>
</tr>
</tbody>
</table>

**RESULTS**

All results are calculated using the within group score as the unit of analysis. For all tests, p < .05 was used as the significance level for significance testing.

Hypothesis One predicted agreement between instructors in the Training Group would increase over
time compared with that of instructors in the control group. To test this hypothesis, a Repeated Measures Analysis of Variance with Time as the within-subjects factor and Training Group as the between-subjects factor was conducted with training condition predicting graded scores at each time index. Results indicated that the cubic trend for both Time frame and the interaction between Time and Training Group emerged as significant predictors of the model, \( F(1, 12) = 5.05, p < .05, \) partial \( \eta^2 = .64 \) and \( F(1, 12) = 10.56, p < .01, \) partial \( \eta^2 = .47 \) respectively. Therefore, data were consistent with

**Graph 1: Standardized grade dispersion by group over time**
the hypothesis. (See Table 1 for means and standard deviations. See also Graph 1 for standardized grade dispersions by group over time.)

**Table 5**

*Self-reported perceived normative behavior indexes of each group over time*

<table>
<thead>
<tr>
<th></th>
<th>Training Group</th>
<th>Control Group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>sd</td>
</tr>
<tr>
<td>Time 1</td>
<td>3.86</td>
<td>1.43</td>
</tr>
<tr>
<td>Time 2</td>
<td>5.07</td>
<td>.85</td>
</tr>
<tr>
<td>Time 3</td>
<td>5.57</td>
<td>1.06</td>
</tr>
<tr>
<td>Time 4</td>
<td>5.89</td>
<td>.63</td>
</tr>
</tbody>
</table>

Hypothesis Two predicted that participants in the training group would report greater perception of agreement with other instructors as time went by when compared with the non-trained group. In order to test this hypothesis, a Two-way Analysis of Variance was employed to analyze whether time and condition interacted to predict normative behavioral perceptions. The data were consistent with the hypothesis, $F(7, 56) = 4.19, p < .01$. (See Table 5 for means and standard deviations for each group over time.) Post hoc analyses indicated individual perceptions of agreement with other instructors at Time One was significantly different from all other times, as were perceptions at Time 4, with scores at Time 4 higher and displaying a general increasing trend across time. Time also emerged as a significant predictor of perceived normative behavior, $F(3,
56) = 6.18, $p < .01$, showing overall improvement for all instructors over time. Therefore, the improvement was significantly greater for the training group compared to the control group, which only exhibited a marginal increase over time. No other unanticipated effects emerged as significant. (See Table 5 and Graph 2 for perceptions of normative behavior over time.)

Hypothesis Three predicted that regardless of condition, perceived normative behavior would be positively associated with self-efficacy of grading. As instructors perceived they were grading speeches in a manner consistent with other instructors, they were predicted to report greater efficacy in their own grading. The scores for all time periods were averaged to get a global index.
for normative behavior and self-efficacy. (For more information on means and standard deviations for self-efficacy over time, see Table 6 and Graph 3.) In order to

Table 6  
Reported self-efficacy of training and control group over time (Self-efficacy index)  

<table>
<thead>
<tr>
<th></th>
<th>Training Group</th>
<th></th>
<th>Control Group</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>sd</td>
<td>Mean</td>
<td>sd</td>
</tr>
<tr>
<td>Time 1</td>
<td>5.78</td>
<td>.87</td>
<td>6.32</td>
<td>.72</td>
</tr>
<tr>
<td>Time 2</td>
<td>6.24</td>
<td>.72</td>
<td>6.38</td>
<td>.52</td>
</tr>
<tr>
<td>Time 3</td>
<td>6.46</td>
<td>.48</td>
<td>6.43</td>
<td>.47</td>
</tr>
<tr>
<td>Time 4</td>
<td>6.58</td>
<td>.40</td>
<td>6.51</td>
<td>.35</td>
</tr>
</tbody>
</table>

Graph 3: Self-efficacy by group over time
test this hypothesis, a step-wise multiple regression analysis was employed to predict self-efficacy. To control for training group, condition was entered in the first step of the regression and normative perceived behavior was entered in the second step. The overall model was significant $R = .77$, $F (1, 54) = 7.934$, $p<.001$. Normative behavioral perception did emerge as a significant predictor of the model, $b = .54$, $t(55) = 3.55$, $p<.01$.

**DISCUSSION**

Given the call for instructor training from the field, this study was necessary to establish an empirical foundation for why training instructors continuously is important and how training affects instructor grading over time. Stitt, Simonds, and Hunt (2003) reported that one-time training increased reliability in grading among new graduate assistants. This study shows that regular training provides continued benefits in grading consistency even among instructors with several years of teaching experience.

Results indicated instructors in the training group became more consistent with their speech grades as time went on. The control group fluctuated over time in terms of actual grade consistency, though their dispersion from mean scores in the final time frame was almost identical to those in the initial time frame. The control and training groups also had very similar dispersions in the beginning, but by Time 4, the training group’s deviation from the mean scores was about half that of the control group’s even as their mean scores remained almost identical. In other words, data showed
that instructors in both groups graded similarly for each
time period, but the training group had significantly
less dispersion than the control group in Time 4.

In Hypothesis Two, we predicted instructors who re-
ceived training would report an increase in levels of per-
ceived agreement in the speech grade they gave com-
pared with other course instructors at the university
over time. Results indicated that training did make a
difference over time. Both instructors who were trained
and those who were not trained reported increased per-
ception of agreement with other instructors over time,
but those who were trained showed higher levels of per-
ceived agreement with others in Time 4 compared with
those who were not trained.

Perceived agreement with other instructors there-
fore increased at a faster rate for instructors who re-
ceived training. One criterion for evaluation of general
education courses is consistency in grading, but results
of this study show that not only does training increase
consistency in grading, training also increases percep-
tions of consistency between instructors. Perceptions of
consistency are correlated with self-efficacy, which
shows instructor confidence in their ability to teach (See
Hypothesis Three results.) If part of an evaluation pro-
cedure is asking instructors directly whether they feel
they are on par or meeting the same standards as their
peers, results of this study would indicate instructors
who have been trained would be more likely to respond
affirmatively.

The control group reported slightly higher (though
not significantly higher) levels of perceived agreement
with other instructors in the initial time frame com-
pared with the training group. In later time frames, the
perceptions flip, with the training group reporting higher normative perceptions (See Graph 2). One possible explanation for the control group’s higher level in Time 1 was that they began the semester with slightly more teaching experience overall ($M = 14.86$, $sd = 8.41$) than the instructors in the training group ($M = 7.29$, $sd = 4.54$), though the difference was not significant, $t (12) = -2.10$, $p < .ns$. Also, instructors in the control condition had taught more semesters at this particular university ($M = 7.43$, $sd = 3.10$) than the instructors in the training group ($M = 4.43$, $sd = 2.70$), $t (12) = -1.93$, $p < .ns$. In spite of these advantages of the control group in terms of teaching experience, the training group performed better over time in terms of grading consistency and normative behavioral perception.

This switch in levels of normative perception is important because perceived agreement with other instructors, or the extent to which public speaking instructors think they grade a speech in a manner consistent with their colleagues, is correlated with self-efficacy (See Hypothesis Three results.). Confidence in one’s ability to teach and grade is a desirable goal for instructors. Results of this study indicate that trained instructors increasingly feel as if they are on the same page when it comes to assigning speech grades, and this perception increases as training goes on.

Hypothesis Three predicted that instructors’ increased perception of grading agreement with other instructors is related to increases in self-efficacy. The more instructors thought their speech evaluations were in agreement with other instructors, the greater their levels of self-efficacy were at all time periods.
One way to measure training effectiveness is to study whether it leads to an increase in how well instructors think they do their job as educators, insofar as that role is tied to grading consistently with other instructors is concerned. When instructors think they are consistent with their peers, they also think they are better able to do their jobs as speech evaluators.

**IMPLICATIONS**

The results of this study show that continual training provided benefits in grading consistency over time. Quite clearly, there is value in providing regular training to faculty members, both new and experienced. The subjects of this study were not new instructors; most had several years of experience teaching. It is important to have instructors who grade consistently to allow for comparability across classes, an important component of course standardization required of general education offerings. The question then becomes: at what point does added training stop providing increased reliability? This study was not able to answer this question, and it is listed as one of its limitations in the next section.

The ethical issue of fairness toward students is important to note here. One could ask, “Why bother,” if the average scores of the two groups were essentially the same for each time period. It is precisely because the two groups’ means were similar that the comparison of the grade dispersions could be undertaken. Overall, even if the control group mean was similar to the training group, their individual scores were more diffused. It would not be fair to have students in a class where the
teacher gives higher grades overall, and in another class where the teacher grades harshly, even if the average of the two instructors is the same as the rest of the faculty combined.

The study also showed that increased grading consistency in the training group led to higher levels of perceived normative behavior. Self-efficacy was shown to be related to perceived normative behavior. When one feels one is on the same page as other faculty members, self-confidence increases. As discussed in the literature section, higher levels of confidence lead to less anxiety and to proactive behavior to do one’s job more effectively.

One role of basic course directors is to provide support to instructors so they can do their job well. Continual training helped improve faculty perceptions that they were grading in a similar way with others, and this was related to higher levels of self-confidence. Training in this group setting may also help the basic course director manage time more efficiently, compared to one-on-one follow-ups with individual faculty. The director could stagger meeting times to accommodate faculty schedules so that each instructor could attend at least a few of the regular meetings. Behnke and Sawyer (1997) suggested that regular meetings between course directors and instructors could increase comparability. Group training meetings like that undertaken in this study is more time-efficient than one-on-one meetings, and provides the added benefit of increased self-efficacy and normative behavioral perceptions.

It is quite plausible that basic course directors might design alternative continuous training modes given the difficulty of finding common meeting times for instructors. For example, instructors might be asked to partici-
pate in an online course that would allow them to evaluate student videos and then view other faculty members’ grades and feedback. An online discussion forum might be set up to support this training and provide an alternative to face-to-face group conversations. It would be useful to assess whether this alternative mode would generate the same training results, not only in terms of increased grading reliability, but also in terms of perceived normative behaviors and self-efficacy.

**LIMITATIONS**

One limitation of the study was the quasi-experimental nature of the design. Faculty were allowed to self-select into control and treatment group, based on their availability to attend sessions. On one hand, this allowed us to have a quasi-control group to compare the training group against. On the other, we had to look closely at the comparability of the two groups. We therefore compared mean scores in the four time periods, and they remained similar for both groups. What changed was the dispersion or the consistency of grades in the treatment group. This is important because it showed that the training did not lead to grade inflation or deflation overall, but it led to a tighter set of scores around the mean for the training group. In other words, their grades became more similar to each other. The control group scores, while averaging the same, had wider fluctuations and therefore, less consistency with each other.

Another limitation is that the study stopped at four time periods. It is quite conceivable that there would be a point of diminishing returns when the reliability...
gained would not be worth the extra effort and resources to bring faculty members together for training. This study was not able to answer that question, but it is worthwhile for future studies to establish this point.

**CONCLUSION**

Universities and colleges with basic courses undergo a great deal of assessment and need to demonstrate consistency across general education instructors. This study showed that longitudinal training over the course of one semester can help improve grading consistency among Public Speaking faculty. Regular training provides continued benefits that may not be achieved in one training session during the beginning of the semester, and also proves beneficial even to instructors who have several years’ teaching experience. Moreover, trained instructors showed higher levels of perceived normative behavior, which is correlated with higher levels of confidence that they can do their job well. As institutions have had to hire adjuncts and graduate teaching assistants to teach basic courses, it is worthwhile to invest resources to provide continual training sessions to help reduce the gap between experienced and inexperienced teachers more quickly, thus helping institutions achieve increased standardization in their basic course offerings.
REFERENCES


Instructor Grade Norming


Turman, P.D., & Barton, M.H. (2004). Bias in the evaluation process: Influences of speaker order,
