Student Progress? Prove It!

Donna Surges Tatum, Ph.D.

Course Goals

Many business and professional people recognize the importance of being able to communicate publicly, because they seek training to improve their skills. Effective communication skills are a highly desired commodity in today's job market. Corporations value such things as team-building, accountability, customer service, total quality management, and 360-degree employee evaluations. That, and the increasingly rapid changes in the workplace, make management acutely aware of the importance of competent communicators. The seas of change are best navigated by those who know how to ask for and give directions.

Butler University responds to this need by offering Public Speaking courses. The purpose of this assessment project is to determine the efficacy of the training Butler provides its students. Careful research design and precise measurement provide the basis for this report.

Demonstrable results in the following areas are the teaching goals of the course:

To enhance delivery skills

To teach methods of organization and critical thinking skills

To increase confidence.

Research Questions

1. Is the evaluation form valid and reliable?
2. Are student raters reliable and consistent when rating their peers?
3. Do students improve their public speaking skills when they take Public Speaking classes?
4. Is inconsistency as a rater related to that person's public speaking ability?
5. Is rater severity related to public speaking ability?

Data Description

The data were collected Spring semester of 1997, from a variety of classes taught by four instructors. One hundred forty-eight students gave 381 speeches which were evaluated by 151 raters using a 29-item, six-point scale instrument. A total of 4925 rating forms are in the database.

Assessment Issues

The assessment of oral communication skills has long been fraught with problems other areas such as math and English do not have. One can administer a test in arithmetic, count the correct answers, compare standardized scores, and come up with a reasonable estimate of a student's ability. The expectations for ability are grade- and age-related, and a common frame of reference has been established over the years.

The communication field is now developing such a clear-cut method of evaluation. This assessment project is using the Meaningful Measurement system which uses the Linacre FACETS extension of the Rasch model as the basis for calculations. It is a method which takes subjective, qualitative observations, and transforms them into objective, quantitative measures. The Meaningful Measurement system is designed to maximize the science of assessment. All raters evaluate four videotaped speeches. This provides common ratings to link and calibrate the raters at this school and others across the country. The rating items are checked for fit and calibration.

The following questions are the psychometric and fairness issues of any situation where raters assess skills.

1. What are appropriate expectations?

What proficiency should be required of a ninth grader, a community college student, or a graduating college senior? Do we know the hierarchy of skills? Have we calibrated the competencies? Do we know which skills should be accomplished at what level and in what order? Our intuition and experience must be backed up with the facts of measurement. The Meaningful Measurement system gives this information to the faculty of Butler University so they can make the proper pedagogical decisions.

2. Are the evaluation instruments sound?

Do the items cover the range of the variable? That is, are there some items that are easier than others? It is not useful if items are bunched together. That would be like giving a test of only simple addition problems. We would not find out the student's true ability, only whether he or she can add. If there is a range of easier to harder items, we can pinpoint with greater accuracy the level of a student's competency.

Do all of the items "fit"? Do they measure what they are intended to? Which items need to be rewritten or dropped? Checking for fit also allows us to be sure we are only measuring one thing at a time, and not confusing issues. (For instance, a story problem on a math test may be more of a reading than math question.) If we are not careful, and try to compare apples to oranges, what we end up with is fruit salad.

The rating form used for this assessment project passed all tests with flying colors. It has 29 items targeted to essential competencies and covers a range of about 90 measure units. The two misfitting items are visual aid quality and use. This is due to the visibility in the classroom, which depends on where the rater is sitting.

ANATOMY OF ASSESSMENT

3. How are differences in raters accommodated? How do we achieve objectivity?

Assessing oral communication skills most often is done by a teacher, or other trained judges, using a rating scale. We know that we all live in our own perceptual world, and attend to different things. Thus, no matter how hard we try for "inter-rater reliability," we will never achieve the ideal of all raters being equal. Instead of a false assumption of sameness, we must address the issue of differences. The most important factor in rating is the consistency with which the judge uses the evaluation form.

When assessing skills, we must be very careful to ensure objectivity in a situation which is subjective by nature. We must have a mechanism to control for levels of severity as well as bias. Meaningful Measurement adjusts for the variations in severity, and flags an inconsistent or biased rater.

4. How can we compare results?

What does a raw score of "65" mean? For example, students are assessed on a 20-item, 4-point rating scale instrument by several different raters. The next year new students are evaluated by some of the old and some new raters. Can we compare the students to each other? One judge is very easy, and gives high ratings. Are those students' raw scores "worth" as much as the raw scores received by students who were rated by a tough judge? How do you come up with a fair ranking? Are the students this year truly better than the ones last year? How do we know for sure?

Meaningful Measurement calibrates all speakers on the same "ruler." This makes it possible to directly compare students from speech to speech, class to class, or year to year.

5. How does a teacher maintain a stable frame of reference throughout the course?

It is difficult to think back to the beginning of the semester, and pull up an accurate recollection of a student's performance. We usually have a general impression, and perhaps a remembrance of a specific skill or two. Referring back to rating forms may help, but it is tedious and fuzzy.

With Meaningful Measurement a teacher can refer to calibrated measures and know precisely how much improvement has (or hasn't) taken place over the semester.

Results

Units of Measure

When reading Meaningful Measurement reports, all numbers are directly comparable. For example, money is in common units; we all know there are 100 pennies in a dollar and that a "dollar" is a "dollar." A dollar is comparable from year to year. We have a common frame of reference. When Dad reminisces about paying 17 cents for a gallon of gas thirty years ago, we know we're paying about ten times that amount today. We can adjust for inflation to determine what the real differences are, yet still be in the same units of measure. When we go to the grocery store to buy food, then to a restaurant for a meal, the bills are both in dollar units. We can compare the price of the ingredients in a tossed salad with what it costs to buy one at a fancy cafe. Even though the situations are different, we can maintain a common frame of reference for the relative costs.

The same situation applies to assessment. When our reports are given, they are in units of measure called "logits." Each logit can have 100 points and has the same properties as a dollar. We can compare one "logit/price" to another. We can add and subtract with logits. Student A's first speech measure is 10.05, and her second measure is 11.45. We know she has progressed by 1.40 logits, or 140 points.

The scale has been calibrated so the origin, or balance point, is "10.00." That means a speech which is of average ability, or a rater who is of average severity, has a measure of 10.00. The lower the number, the less able or less severe a person is measured. Measures higher than 10.00 indicate more ability or severity than that of the "average" speaker or rater.

We have established and maintained a metric that can be used from year to year, and situation to situation. We have the means to track and assess improvement.

Raters

The 151 raters are examined to determine how consistent they are when rating speeches. An investigation of the fit statistics shows that 84% of all raters are "good." That is, they are internally consistent and are able to maintain a stable frame of reference when evaluating speakers. This means we can trust the speech measures. The raters are not behaving erratically.

The raters' mean severity measure is 10.00. They fit well, but cover a wide range of severity from easy to hard when rating speeches.

Items

The Item Map below shows the hierarchy of items. The Butler University speech communication faculty determined that these are the essential competencies required of the students when giving a speech.

The calibration of the items goes from easy to hard. The lower the number, the easier the item is to accomplish. The items cover a range of 95 points. The point biserials show that all the items are related, and define a common variable. The separation reliability is .99.

At Level 1 the easist thing for the students to do is to show their knowledge/mastery of the topic, pick a worthy topic, and appear trustworthy.

At Level 2 the next easiest items include showing the relevance of the topic, using appropriate language, being understandable, using materials appropriate to the audience, limiting the topic, and using clear language.

At Level 3 the visual impression of the speaker, word choice and establishing common ground are a bit more difficult. A well-organized speech using good quality support are next in the hierarchy.

At Level 4 ethical and appropriate emotion appeals are slightly above average in difficulty, as are eye contact and a poised demeanor.

At Level 5 a conversational style and variety in vocal delivery are more difficult to accomplish. The quality and use of visual aids are also in this strata.

It is progressively more difficult to use a sufficient quantity of verbal support with a variety of sources, and to respond to audience feedback. Well-presented support with citations and establishing a context is harder to do.

At Level 6 an enthusiastic delivery is quite difficult on this scale. The flow of the speech with preview/review, signposting, and transitions is also at this point.

Finally, **at Level 7** fluency and smoothness in vocal delivery is the second most difficult thing for a speaker to do. Gestures are the hardest for a speaker to effectively accomplish at Level 8.

Speech Results

Ninety-four students in the basic course gave at least two prepared presentations, 88 gave three, and 11 gave four. Thirty-two students in the advanced course gave two prepared presentations.

The mean of all speeches is 11.64, or 164 points above the mythical average speaker at 10.00. This shows the Butler University student body is an accomplished group. The separation of 8.18 and standard deviation of .75 demonstrate there is a wide range of ability in this sample. The normal, bell-shaped distribution shows speakers' ability from about 8.20 to 13.60, a range of over 500 points.

ITEM MAP

EASY	SPEAKER	MESSAGE	AUDIENCE
1	mastery trustworthy	worthy topic	
2	understandable	appropriate language limit topic clear language	relevance materials appropriate
3	visual impression word choice	well-organized	common ground
4	eye contact demeanor		ethical emotion appropriate emotion
5	conversational variety	aid quality aid use quantity support	responds to feedback
6	enthusiastic	well-presented support flow of speech	
7	fluency		
8	gestures		
HARD			

Speaker Improvement - 2 Speeches

Ninety-four students gave two prepared presentations. The mean measure for the first speech is 11.17. The second speech measure averages 11.45. This is an average gain of over a quarter of a logit, or 28 points.

A paired samples t-test tests the hypothesis of whether the first round of speeches is the same as the second round of speeches.

In other words, does training make a difference? Do speakers improve? The answer is "Yes!"

The t-value of 4.56 with a significance of .000 means we are absolutely sure: The two groups are truly different, and the improvement is not due to chance.

Speaker Improvement - 3 Speeches

We know students significantly improve from their first to their second speeches. Now we want to know if they continue to gain in ability.

Learning does not stop after two rounds of speeches. Students have not learned all there is to know about public speaking after just two speeches, for they continue to improve as shown by the following table.

Seventy-seven students gave three prepared presentations. The results of this group are shown, for instance, through the paired samples t-test of the second and third round of speeches.

The mean of this group of second speeches is 11.49, and the mean of the third is 11.71. Again the students improved — this time by .22 logits, or 22 points.

The significance of .000 means we are 100% sure the third round of speeches is truly different from the second round.

Speaker Improvement - 4 Speeches

Eleven students gave a fourth speech. These students improved another 30 points. The t-value of 2.33 with a significance of .045 means we are 95.5% sure that the fourth round gain is due to training.

Speaker Improvement - Advanced Class

Thirty-two students in the advanced classes gave two prepared presentations. These students continue to improve by 35 points. (In reality this is the fourth and fifth speeches for these students because they already had the basic course.) The t-value of 4.08 with a significance of .000 means we are absolutely sure the advanced training has an effect.

Rater Consistency and Speaker Ability

A Mean square (MNSQ) fit statistic evaluates the consistency of the rater. A mean square of 1.0 is exactly what is expected; .7 to 1.3 is normal. But a mean square of 1.5 means there is 50% more "noise" in a rater's evaluations, and 1.9 90% more variance than expected.

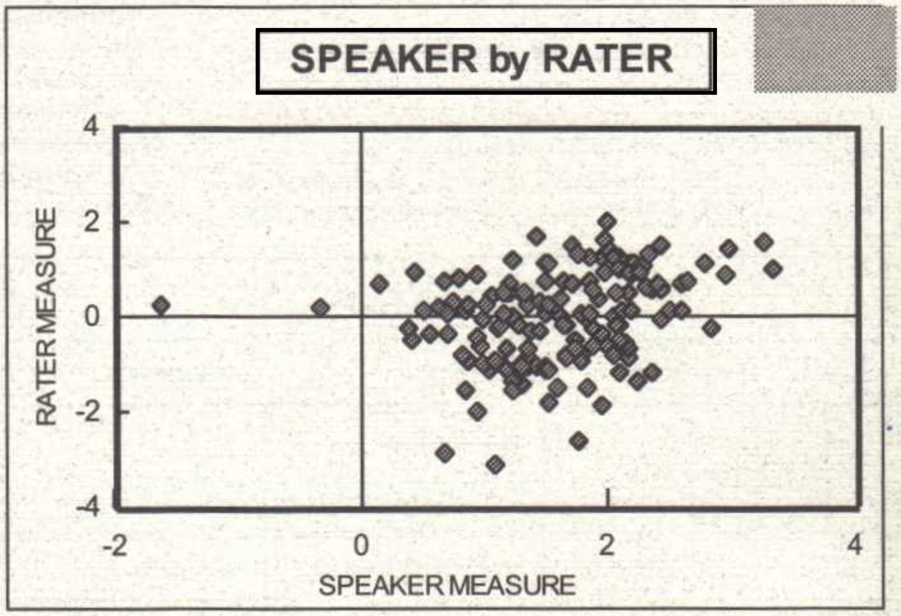
A rule of thumb is to look closely at any response pattern with a mean square of more than 1.4, or a standardized fit over 2. When this occurs, a red flag waves in the researcher's mind, and a close examination of the data is warranted to determine the cause of the misfit. It may be that the rater is consistently inconsistent and should not be used for assessment purposes, or perhaps the rater had a bad day.

Some raters have mean squares and fits that are almost too quiet, mean squares of .5 or below. They are close to Guttman-like in their consistency. Their evaluations hold no surprises or randomness. They are rating holistically instead of discriminating among the items.

Fifteen of the 152 raters are inconsistent, and 10 are overly consistent. The table above shows these 25 rater fit statistics with their speech measures. But there is no relationship between a rater's consistency and speech ability.

Rater Severity and Speaker Ability

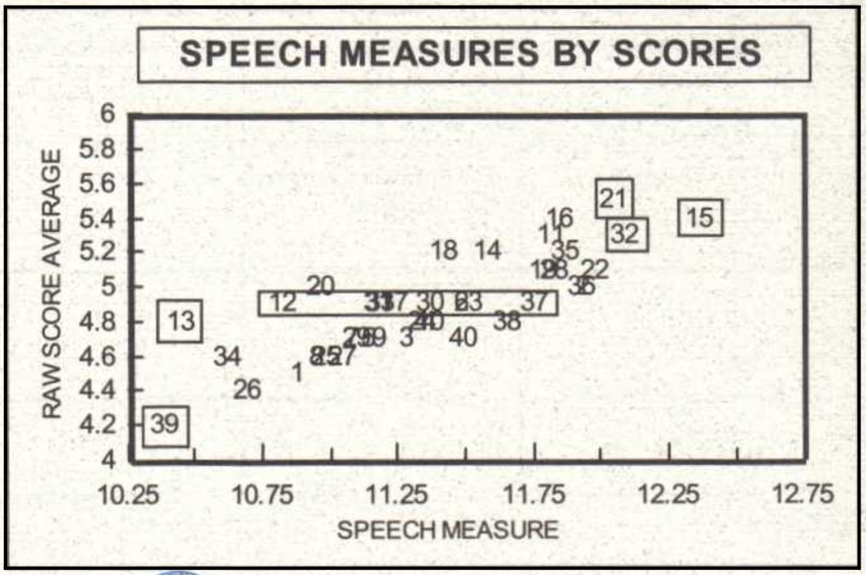
The graph below shows there is not a clear relationship between a person's severity as a rater and their ability as a speaker. Some excellent speakers are easy raters, and some poor speakers are quite severe.



Measures and Raw Scores

The next graph demonstrates the importance of objective measures rather than a proportion of raw scores. When the severity of the rater is taken into consideration, the results can be different.

Forty speeches were randomly chosen from the database. The average of the raw scores is plotted against the speech



measure. Eight speakers have a raw score of 4.9. However, their measures range from 10.82 to 11.75, a difference of 93 points.

The worst speech is #39 with a raw score of 4.2 and a measure of 10.39, yet the second lowest speech, #13, has a measure of 10.45 and a raw score of 4.8.

Speech #21 has the highest raw score, 5.5, but is third in ability after the raw scores are conditioned into measures (behind #32 with 5.3 and 12.09, and #15 at 5.4 and 12.36).

Now we have a method to not only ensure, but prove fairness in the judging process. This is extremely important in grading and other high-stakes assessments.

Discussion

Meaningful Measurement Results

The results show that training in Public Speaking produces positive results. Students significantly improve from their first to second speeches, and they continue to do so in subsequent speeches and in subsequent advanced classes.

We can have confidence that these outcomes are not dependent upon a particular teacher, because the students came from eight classes taught by four different teachers. The Butler University Speech Department is fulfilling its mission, and should be commended for the excellent job it is doing in training its students.

This study also demonstrates:

1. Students are useful, reliable raters. Since audience analysis is taught as an important factor when preparing a speech, we can now derive speech measures from the entire class instead of only one grade from one teacher.

2. Averaging raw scores does not produce reliable speech measures.

3. A student's consistency as a rater is unrelated to his or her ability as a speaker.

4. A student's severity as a rater is unrelated to his or her ability as a speaker.

5. The hierarchy of item difficulty improves our concept of what is required for public speaking ability. Now it is possible to identify the items that turn a poor speaker into a good one. Expectations for progress can be realistic and predictable. Teaching methods improve because information can be sequenced according to actual student development.