

Appendix 3

Reliability, Validity and Bias Testing Report: The Competent Speaker Evaluation Form

APPENDIX C

RELIABILITY, VALIDITY, AND BIAS TESTING

Traditional Analysis

The following section describes the testing processes from the original study using traditional statistical analyses to determine the reliability and validity of **The Competent Speaker Evaluation Form**. Further investigation for this second edition then is reported in the Rasch Analysis section that follows.

Development of a Testing Videotape

In order to test the reliability and validity of **The Competent Speaker Evaluation Form** and criteria, a videotape was developed with student speeches in an actual classroom environment at a Midwestern University. The student speeches were informative presentations lasting approximately five minutes in duration. A group of six Graduate Teaching Assistants (GTAs) selected approximately 40 videotaped classroom speeches from the previous semester's presentations. Utilizing the form and criteria, the GTAs rated these speeches as either unsatisfactory, satisfactory, or excellent. From this initial pool of 40, the public speaking course director and one of the GTAs selected 12 student speeches, four at each level of competency. These 12 speeches were transferred to one master tape for training and rating purposes. The student sample represented on the rating video was mixed by gender (five females and seven males) and by ethnicity (nine Whites, one Black, one Hispanic, and one Filipino). The 12 speeches then were randomly ordered 12 different times, utilizing a table of random numbers, and placed on tapes that were sent to 12 speech communication professionals at 12 different universities. Additionally, one speech exemplary of each level of competency was selected as an anchor by which raters could become familiar with performance at each level of competency before rating the sample of 12 speeches. The three anchor speeches were placed at the beginning of the master tape and were identified as to the level of competency each represented.

Raters

The 12 raters were speech communication professionals teaching at colleges and universities in the U.S. Nine of the raters held a Ph.D. or equivalent, while three of the raters held master's degrees. The raters' experience in teaching ranged from 4 years to 25 years as estimated by the date of receipt of the raters' terminal degrees. Raters were eight females and four males, eleven of which were Anglo and one Hispanic.

Raters received a packet containing instructions for self-training on the use of the speech evaluation form and criteria and the tape with the 12 student presentations. Specifically, the raters were instructed to: (a) review the standards and criteria for the competencies before viewing any speeches; (b) view the three exemplary speeches while simultaneously reviewing the standards and criteria; and (c) view each of the 12 speeches without making any formal evaluation, review the standards and criteria as they pertained to that speech, and finally, view the speech one more time and enter the evaluation on the rating form.

Overall inter-rater reliability for the students' total score on the instrument was high for the 12 raters with Ebel's (1951) coefficient reading .92. Inter-rater reliability was also examined for each of the eight competencies. The 12 raters achieved a high degree of reliability on the eight competencies with Ebel's coefficient ranging from .90 to .94.

In addition to using 12 speech communication professionals to test reliability, other reliability testing was conducted utilizing 10 GTAs as raters. The raters were from two Midwestern universities. Half of the GTAs had utilized the speech evaluation form for one semester and the other half were given a brief training in the instrument's use. An inter-rater reliability test for the GTAs generated a Cronbach coefficient of .76.

In addition to the GTAs, inter-rater reliability testing was conducted with a small pool of community college speech instructors (N=3). They received a brief training with the instrument before evaluating the 12 videotaped speeches. This inter-rater reliability test generated a Cronbach coefficient of .84.

26 The Competent Speaker Speech Evaluation Form

Validity

In addition to reliability testing of **The Competent Speaker** form, the instrument and its criteria currently meet content or face validity. That validity can be argued based on the extensive literature review conducted during the process of development of the instrument by the 11 member subcommittee of the NCA Committee on Assessment and Testing (see Appendices A and B of this manual).

Also regarding the validity of the instrument, two studies testing for convergent validity have been conducted. First, a correlation of scores on the public speaking portion of the Personal Report of Communication Apprehension (McCroskey, 1970) and scores derived using **The Competent Speaker** form in an introductory speech class indicated inverse directional convergent validity. That is, as speech scores using **The Competent Speaker** form increased, scores on the public speaking items of the PRCA decreased. Second, a correlation of scores on the seven public speaking items of the Communication Competency Assessment Instrument (Rubin, 1982), derived from entrance and exit interviews with students in an introductory speech class, were correlated with **The Competent Speaker** scores from the same class. Positive directional convergent validity for the two instruments was indicated: scores on speeches, rated using **The Competent Speaker** and scores on the public speaking items of the CCAI both increased.

Cultural Diversity

In addition to **The Competent Speaker Evaluation Form** and criteria undergoing reliability and validity scrutiny, several other tests evaluated the form in regard to ethnic and gender bias. One study compared the 12 speech communication professionals' ratings of 12 speeches to the ratings of the same speeches by a sample of 28 minority students using the speech evaluation form. An inter-rater reliability test of the minority students as a group generated a Cronbach's alpha coefficient of .76. When combined with the 12 speech communication professionals, the coefficient remained .76. In another two-pronged diversity study, actual speech evaluations in the classroom (N=260) were examined both by ethnicity and by gender. In an analysis of variance, no significant difference was found in the ratings of speeches of White (m=86.46), Black (m=82.92), Hispanic (m=85.25), or Asian (82.33) students (F=.16). Additionally, there was no significant difference in speech scores of female (m=86.47) or male (m=85.70) students using the speech evaluation form (F=.41).

Normative Data

Although normative data are provided for training purposes with the videotaped anchor speeches, **The Competent Speaker Evaluation Form** and standards are criterion referenced. The competencies and criteria were developed based upon the literature investigated.

Conclusions of Traditional Analysis

Initial and subsequent testing of **The Competent Speaker Evaluation Form** indicates that the instrument is psychometrically sound in terms of reliability and validity.

As described in the following section, larger and more diverse pools of raters and greater diversity of student populations are evaluated for this edition of **The Competent Speaker Evaluation Form**. These studies are found to be confirmatory of the original results.

Rasch Analysis

After researchers develop an instrument expected to measure a phenomenon, they test it in the field. Data analysis of the collected sample(s) determines whether or not the instrument is deemed a valid mode of measurement.

In 1953 Georg Rasch, a Danish mathematician, developed a unique model for item analysis. This method produces results that are distinctly different from traditional statistical analysis. A statistical analysis describes a one-time event. The elements of the event are inextricably bound together into one observation. Those elements are, in this case, the items on the evaluation form, the raters using it, and the speeches they are judging.

The results of a traditional statistical analysis are not generalizable or comparable across samples or time. However, the unique feature of the Rasch model is that it allows the researcher to separate the elements under investiga-

The Competent Speaker Speech Evaluation Form 27

Source: Morreale, S., Moore, M., Surges-Tatum, D., & Webster, L. (2007) *The competent speaker speech evaluation form* (2nd ed.). National Communication Association. pp. 26-27.