Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 3, Number 3, November, 1992 ISSN=1531-7714

Reducing Errors Due to the Use of Judges

Lawrence M. Rudner
ERIC Clearinghouse on Assessment and Evaluation

Many alternative forms of assessment--portfolios, oral examinations, open-ended questions, essays--rely heavily on multiple raters, or judges. Multiple raters can improve reliability just as multiple test items can improve the reliability of standardized tests. Choosing and training good judges and using various statistical techniques can further improve the reliability and accuracy of instruments that depend on the use of raters.

After identifying several common sources of rating errors, this article examines how the impact of rating errors can be reduced.

UNDERSTANDING RATING ERRORS

There are numerous threats to the validity of scores based on ratings. People being rated may not be performing in their usual manner. The situation or task may not elicit typical behavior. Or the raters may be unintentionally distorting the results. Some of the rater effects that have been identified and studied are:

- The halo effect. The impressions that an evaluator forms about an individual on one dimension can influence his or her impressions of that person on other dimensions. Nisbett and Wilson (1977), for example, made two videotapes of the same professor. In the one, the professor acted in a friendly manner. In the second, the professor behaved arrogantly. Students watching the friendly tape rated the professor more favorably on other traits, including physical appearance and mannerisms.
- Stereotyping. The impressions that an evaluator forms about an entire group can alter his or her impressions about a group member. In other words, a principal might find a mathematics teacher to be precise because all mathematics teachers are supposed to be precise.
- Perception differences. The viewpoints and past experiences of an evaluator can affect how he or she interprets behavior. In a classic study, Dearborn and Simon (1958) asked business executives to identify the major problem described in a detailed case study. The executives tended to view the problem in terms of their own departmental functions.
- Leniency/stringency error. When a rater doesn't have enough knowledge to make an objective rating, he or she may compensate by giving scores that are systematically higher or lower.
- Scale shrinking. Some judges will not use the end of any scale.

MINIMIZING RATING ERRORS THROUGH TRAINING

An established body of literature shows that training can minimize rater effects. In 1975, Latham, Wexley, and Purcell used training to reduce rater effects among employment interviewers. Since then, a variety of training programs have been developed in both interviewing and performance appraisal contexts.

For example, Jaeger and Busch (1984) used a simulation to train judges in a three-stage standard-setting operation. After working through the simulation, the judges clearly understood their rating task.

Pulakos (1986) trained raters in what types of data to focus on, how to interpret the data, and how to use the data in formulating judgments. This training yielded more reliable (higher inter-rater agreement) and accurate (valid) ratings than no training or "incongruent" training (training not tailored to the demands of the rating task).

This literature suggests that rater training programs should:

- familiarize judges with the measures that they will be working with,
- ensure that judges understand the sequence of operations that they must perform, and
- explain how the judges should interpret any normative data that they are given.

CHOOSING JUDGES

The choice of judges may have a significant influence on scores. Hambleton and Powell (1983) have done an excellent job of identifying many of the issues involved in choosing judges. Their recommendations to some common questions are:

- **Should demographic variables be considered when selecting judges?** Hambleton and Powell argue that demographic variables such as race, sex, age, education, occupation, specialty, and willingness to participate should be considered in the selection of judges. The composition of the review panel often lends credibility to the overall effort.
- **Should expert judges be preferred to representatives from interest groups?** The authors suggest that, whenever possible, review panels should be composed of both experts and representatives from **interest groups.**
- **Should the review panel split into separate working groups?** The authors argue that smaller working groups should be formed when the review panel is too large to permit effective discussion and when the ratings are going to be compared across groups to assess reliability or to cross check validity.

USING STATISTICAL TECHNIQUES

The difference between a rater's average and the average of all ratings is called the "rater effect." If the rater effect is zero, no systematic bias exists in the scores. Because of rater errors such as those discussed earlier, the rater effect is rarely zero.

If all the judges rate everyone being evaluated, some rater effects may not be a problem: The candidates all realize the same benefit or penalty from the rater's leniency or harshness. The ranks are not biased, and no one receives preferential treatment.

However, an issue arises if different sets of multiple raters are used--a common situation when scoring essays, accrediting institutions, and evaluating teacher performance. Candidates evaluated by different sets of multiple raters may receive biased scores because they drew relatively lenient or relatively harsh judges.

Several approaches may be followed to adjust potentially biased ratings given by different sets of multiple raters. Compared with simply averaging each candidate's ratings--in other words, doing nothing--these statistical approaches have been shown to reduce measurement error and increase accuracy. When applied to actual performance data, they typically produce substantial adjustments and change significant numbers of pass/fail decisions.

Three statistical approaches discussed in the literature are (see Houston and Svec, 1991):

- ordinary least squares regression, where the observed rating is viewed as the sum of the candidate's true ability, a rater effect, and random error;
- weighted least squares regression, where each rater's score is weighted by a measure of the rater's consistency; and
- imputation of missing data, where actual data are used to estimate scores for the candidates that the rater did not evaluate.

The imputation approach is most appropriate when each rater evaluates only a few candidates. The weighted regression approach is most appropriate when variations are expected in rater reliability.

REFERENCES

Dearborn, D.C., and H.A. Simon. (1958). Selective perception: A note on the departmental identification of executives, *Sociometry*, June, 140-148.

Hambleton, R.K., and S. Powell. (1983). A framework for viewing the process of standard setting. *Evaluation and the Health Professions*, 6(1), 3-24.

Houston, W.M., M.R. Raymond, and J.C. Svec. (1991). Adjustments for Rater Effects. *Applied Psychological Measurement,* 15(4), 409-421.

Jaeger, R.M., and J.C. Busch. (1984). The effects of a Delphi modification of the Angoff-Jaeger standard setting procedure on standards recommended for the National Teacher Examination. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Nisbett, R.E., and T.D. Wilson. (1977). The halo effect: Evidence for the unconscious alteration of judgments. *Journal of Personality and Social Psychology*, 35, 450-456.

Pulakos, E.D. (1986). The development of training programs to increase accuracy on different rating forms. *Organizational Behavior and Human Decision Processes*, 38, 76-91.

Descriptors: *Error of Measurement; Evaluation Methods; *Evaluators; *Interrater Reliability; Least Squares Statistics; Rating Scales; Regression (Statistics); Scaling; Scores; *Scoring; Test Interpretation; *Training; Validity

Citation: Rudner, Lawrence M. (1992). Reducing errors due to the use of judges. *Practical Assessment, Research & Evaluation*, 3(3). Available online: http://PAREonline.net/getvn.asp?v=3&n=3.