DOCUMENT RESUME

ED 388 688 TM 023 916

AUTHOR Livingston, Samuel A.; Sims-Gunzenhauser, Alice
TITLE Who Will Watch the Watchers? Setting Standards for Classroom Observers.
PUB DATE 22 Apr 95
NOTE 8p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22, 1995).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Beginning Teachers; *Classroom Observation Techniques; Documentation; Elementary Secondary Education; Higher Education; *Interrater Reliability; *Judges; Standards; *Test Construction; Test Reliability
IDENTIFIERS Accuracy; Performance Based Evaluation; *Praxis Series; *Standard Setting

ABSTRACT

A study was conducted to provide information for setting two separate standards, the accuracy score and the documentation score, for the Praxis III: Classroom Performance Assessment (Praxis III). Praxis III is intended for making instructional and licensing decisions about beginning teachers. This standard-setting study was a person-judgment study. Test-takers were actual assessor trainees who had taken the Praxis III Assessor Proficiency Test. The judges were five developers of the Praxis III assessment. A test developer who did not serve as a judge selected 15 record of evidence forms completed by assessor trainees as examples of the test-takers' performance to be judged. Although five judges rated the examples, only four were able to meet to discuss the tests and arrive at consensus. Judgments before discussion showed many disagreements among judges. More than 2 hours of discussion were required to reach agreement on 12 test-takers, and consensus was reached by only 3 judges on the remaining 3. Results indicate that under some conditions, there can be great value in trying to get judges to reach consensus. (Contains one table and one figure.) (SLD)

Who Will Watch the Watchers? Setting Standards for Classroom Observers[1]

Samuel A. Livingston
Alice Sims-Gunzenhauser

Educational Testing Service

The Praxis III assessor

The Praxis III: Classroom Performance Assessment is intended for making instructional and licensing decisions about beginning teachers. It consists of a series of classroom observations, each preceded and followed by an interview with the beginning teacher. The person who does the interviewing and observing is called the assessor. After each observation (and the accompanying interviews), the assessor assigns nineteen separate scores to the beginning teacher's performance. Each of these nineteen scores refers to a different criterion, i.e., a particular aspect of the beginning teacher's performance. The scores are expressed on a scale with six levels: 1.0, 1.5, 2.0, 2.5, 3.0, and 3.5. Each of the nineteen scores represents a judgment by the assessor, based on events that occurred in the classroom or in the interviews. The assessor records the scores on a "Record of Evidence" form, which contains spaces for written documentation to support each of these nineteen judgments.

Clearly, the assessor's role is central. The validity of the assessment and the fairness of the decisions based on it depend directly on the competence of the assessors.[2] Therefore, the qualifications and the training of the assessors are critically important. The experienced educators selected to become assessors must complete a five-day assessor training course. They practice interviewing, observing and taking notes on classroom events, identifying the Praxis III criteria relevant to a classroom event, assigning scores on the Praxis III score scale, and providing written documentation. They then conduct an observation (and the accompanying interviews), and complete the Record of Evidence form.

The Assessor Proficiency Test

The assessor's training concludes with the Assessor Proficiency Test. The purpose of the Assessor Proficiency Test is to make sure that every assessor who goes into a classroom to conduct a Praxis III assessment can

---

[1]This paper was presented at the Annual Meeting of the American Educational Research Association, April 22, 1995, in San Francisco, California.

[2]Educational Testing Service recommends that, to the extent possible, each observation of the beginning teacher's performance be conducted by a different assessor. Any decision about a beginning teacher should be based on observations made on at least two different occasions and by at least two different assessors. (See Guidelines for Proper Use of The Praxis Series: Professional Assessments for Beginning Teachers™.)

correctly apply the scoring rules and complete the Record of Evidence form. The Assessor Proficiency Test is based on a videotape of a Praxis III "assessment cycle": the pre-observation interview, the observation of the teacher's performance in the classroom, and the post-observation interview. The assessor trainee taking the test watches the videotape, takes notes, and completes the Record of Evidence form.

Each test-taker receives two scores on the Assessor Proficiency Test: an accuracy score and a documentation score. These scores are expressed on a scale of 0 to 100. The accuracy score measures the extent to which the test-taker has correctly applied the scoring rules. The documentation score is a measure of the quality of the written documentation the test-taker provides.

The accuracy score is computed from the scores that the test-taker assigns to the videotaped performance. For each of the nineteen criteria there is a "juried" score that serves as the correct answer. The accuracy score is based on the differences between the scores assigned by the test-taker and the juried scores.

The documentation score is computed from ratings of the test-taker's documentation. The rater, a specially qualified Praxis III staff member, assigns three separate ratings to the documentation for each of the nineteen criteria. The documentation score is computed from these 57 ratings.

The study

The standard-setting study was a person-judgment study, rather than an item-judgment study. The test-takers were actual assessor trainees who had taken the Assessor Proficiency Test. The study used a contrasting-groups approach. The classification of the test-takers into contrasting groups was based on holistic judgments of their performance on the test itself. The product of this performance was the Record of Evidence forms they completed while taking the test.

The study was intended to provide information for setting two separate standards, one for the accuracy score and one for the documentation score. Therefore, the study included two separate sets of judgments: (1) judgments of the accuracy of the scores that each test-taker assigned to the videotaped performance and (2) judgments of the adequacy of the written documentation the test-taker provided.

The judges

The judges for this study were five developers of the Praxis III assessment. Though their roles in the development process differed, all five judges were thoroughly familiar with the videotaped performance used in the test. These five judges were not selected as a sample from some larger population of possible judges. They were the five individuals best qualified to make the judgments called for by the study.

The procedure

The procedure for the study consisted of the following steps:

1. A Praxis III developer who did not serve as a judge reviewed several Record of Evidence forms completed by assessor trainees taking the Assessor Proficiency Test. She selected a sample of fifteen Record of Evidence forms representing a wide range in the accuracy of the scores assigned and in the quality of the documentation. These were the examples of test-takers' performance to be judged.

2. This same Praxis III developer rated the documentation on each of the fifteen selected Record of Evidence forms, providing the numerical ratings that would later be used to compute the documentation scores.

3. The five judges individually reviewed the fifteen Record of Evidence forms and made two holistic yes-or-no judgments: (1) whether the criterion scores awarded by the test-taker to the videotaped performance were acceptably accurate, and (2) whether the test-taker's written documentation reflected acceptable performance as an assessor. Each of the judges was given the Record of Evidence forms in a different, randomly determined sequence. The judges were not given any information about the numerical ratings that had been assigned to the documentation on these fifteen Record of Evidence forms.

4. Four of the five judges (one judge was unavailable) met to discuss their judgments of the documentation, resolve disagreements, and reach a group consensus judgment of the documentation produced by each individual test-taker.[3]

The portion of the study that involved the documentation scores differed in some important ways from the portion that involved the accuracy scores. It involved a type of performance that is often measured in performance assessments -- creating a written document. It presented the judges with a situation in which they could not know (even approximately) the numerical scores of the performances they were judging. And it offered an opportunity to explore the effects of attempting to resolve differences between the judges. For these reasons, the rest of this paper will focus on the documentation scores and judgments.

The results

The relatively small number of test-takers and judges in the study makes it practical to show the raw data -- the individual scores and judgments -- in

[3]The limited availability of the five judges made it impossible to get consensus judgments for both accuracy and documentation. We gave priority to obtaining consensus on documentation, because much of the judgment required to evaluate accuracy had been a part of the process of determining the juried scores.

a table. Table 1 shows the numerical score assigned to each test-taker's written documentation and the judgments it received from each individual judge (before discussion) and from the group (after discussion). The test-takers are shown in order of their numerical scores; the judges are shown in order of the number of favorable judgments they awarded. The numerical scores range from 54 to 83, with a mean of 68.6 and a standard deviation of 9.1. (For comparison, the scores of 31 assessor trainees taking the Assessor Proficiency Test in one state had a mean of 70.3 and a standard deviation of 13.5; the scores of 30 assessor trainees taking the Assessor Proficiency Test in another state had a mean of 74.4 and a standard deviation of 8.9.)

The individual judgments, made before discussion, showed many disagreements between judges. Even when judges agreed closely as to how many of the test-takers had performed acceptably, they disagreed as to which test-takers had performed acceptably. Predictably, the process of achieving consensus proved to be difficult, even with only four judges participating. Finally, after more than two hours of discussion, the group reached agreement by all four judges on twelve of the fifteen test-takers and by three of the four judges on each of the remaining three test-takers. For three of the fifteen test-takers, the group consensus judgment was different from the majority of the individual judgments made before discussion, and in one of these cases, a single judge presuaded the rest of the group to change their judgments.

The group consensus judgments agreed much more strongly with the numerical scores than the individual judgments did. Only one of the five judges -- Judge B -- made individual judgments that agreed strongly with the numerical scores. The correlations of the numerical scores with the individual judgments were .15, .69, .23, .34, and .23 for the five judges; the correlation of the numerical scores with the group judgments was .74. Yet, the correlation of Judge B's individual judgments with the group judgments was only .50.

In setting a standard to be used for making decisions about individuals, the key question to be answered from a person-judgment study is: "Given the numerical score assigned to a performance, what is the probability that the performance will be judged acceptable?" One statistical procedure commonly used to estimate this kind of relationship is called logistic regression. This procedure assumes that the relationship can be described on a graph by a curve of a particular shape -- the shape of the curves in Figure 1. The data determine the extent to which the curve is shifted left or right and the extent to which it is compressed (giving it a steeper slope) or elongated (giving it a shallower slope).[4]

---

[4]This curve has the mathematical equation

$$P = \frac{1}{1 + e^{-a+bx}},$$

where P is the probability, x is the score, e is the mathematical constant 2.71828..., and a and b are parameters estimated from the data.

Figure 1 shows the logistic regression curves describing the relationship between the numerical scores and the judgments. The horizontal scale represents the numerical score; the vertical scale represents the probability of a favorable judgment. The vertical lines at scores of 54 and. 83 indicate the range of scores of the fifteen test-takers included in the study. Figure 1 shows clearly that the individual judgments made by four of the five judges were weakly related to the numerical scores and that one judge's individual judgments and the group consensus judgments were strongly related to the scores. This relationship can also be seen in Table 2, which shows the estimated probabilities at five selected points on the score scale.

Figure 1 reveals some interesting things about the comparison between the group judgments and the individual judgments. For scores below 68 -- roughly the lower half of the range of scores of the fifteen test-takers in the study -- the probability of a favorable judgment from the group, after discussion, was lower than the probability of a favorable judgment from any individual judge before discussion. Although Judge B's individual judgments agreed with the numerical scores almost as strongly as did the group judgments, Judge B was much more likely to make a favorable judgment, particularly for a test-taker whose score was in the middle of the range.

When this type of analysis is used for setting a standard, the standard-setters often focus on the score for which the probability of a favorable judgment is .50. Above this score, the majority of the judgments tend to be favorable; below this score the majority of the judgments tend to be unfavorable. On the graph, this score is indicated by the point at which the curve crosses the horizontal line for probability .50. This point is not stably estimated when the slope of the curve is shallow, as it is for four of the five judges. If the study had used a different sample of fifteen test-takers, the curve for any of these four judges might well have crossed the .50 line at a very different place. However, the steep slope for the group consensus judgments suggests that this result -- the point at which the probability of a favorable judgment from the group first exceeds .50 -- would tend to be similar if the same judges were to judge another sample of test-takers.

Implications of the results

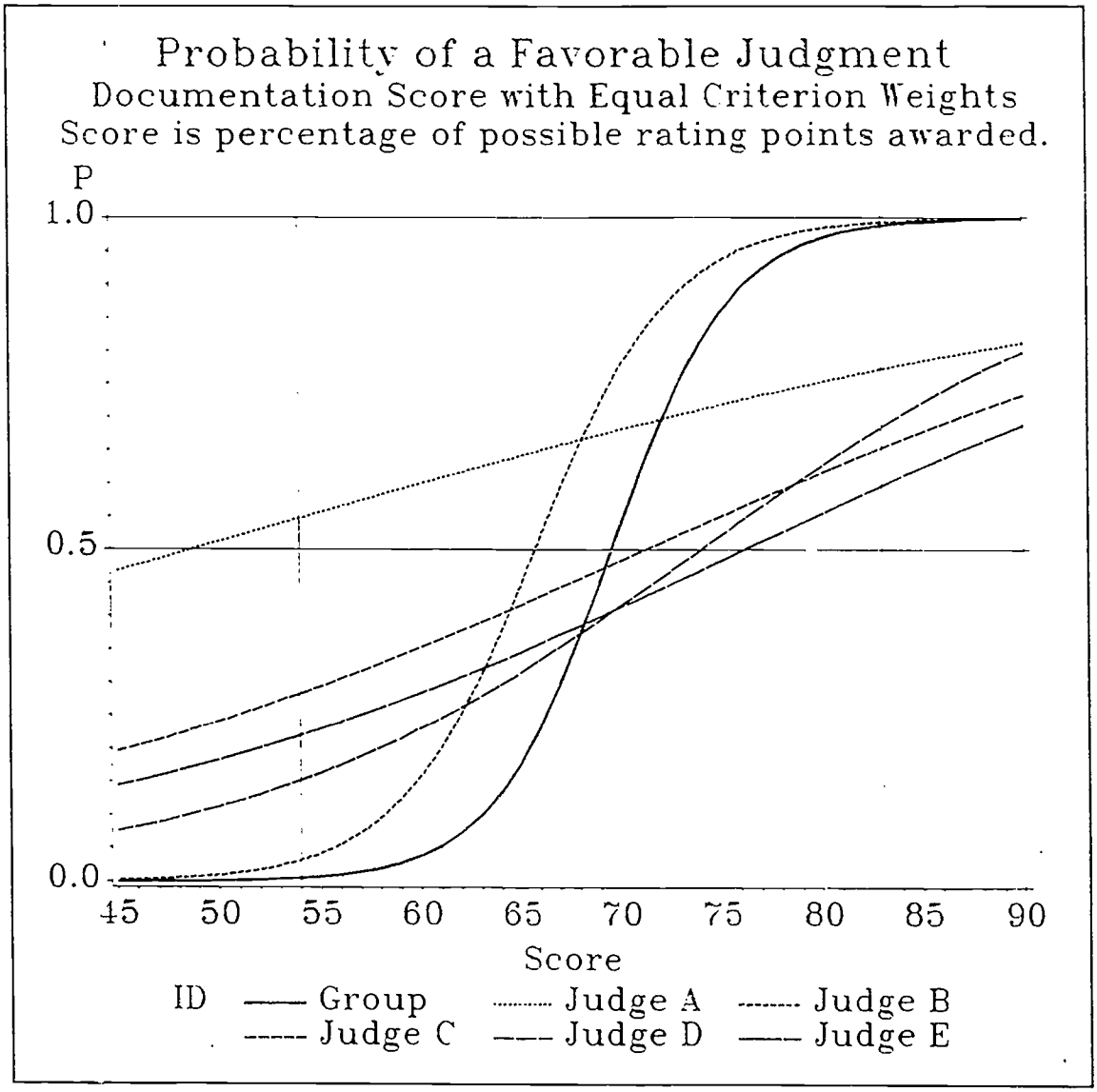
The results of this study have some implications that go beyond the specific test involved. They indicate that, at least under certain conditions, there can be great value in trying to get the judges to reach consensus. The conditions of this study would seem to be particularly favorable for reaching consensus. The number of judges was small, and the four judges who participated in the consensus process were accustomed to working with each other. Despite these favorable conditions, the judges took more than two hours to reach consensus on fifteen test-takers, and in three cases the consensus was not unanimous. The effort proved worthwhile when the group judgments agreed strongly with the numerical scores, providing useful, relevant, believable information for choosing a passing score for the test.

Table 1.
Documentation scores and judgments.

Numerical Score	Individual Judgment by Judge					Favorable judgments	Group consensus[5]
	A	B	C	D	E		
54	0	0	0	0	0	0	0
54	1	0	1	0	0	2	0
61	1	1	0	1	0	3	0
63	1	0	0	1	1	3	0*
63	1	0	1	0	1	3	1**
65	0	0	1	0	1	2	0
67	1	1	0	0	0	2	0
68	0	1	0	0	0	1	0
69	0	0	0	1	0	1	0*
72	1	1	1	0	0	3	1
75	1	1	0	0	1	3	1
78	0	1	0	0	0	1	1
79	1	1	1	1	1	5	1
80	1	1	1	1	0	4	1
83	1	1	1	1	1	5	1
Favorable judgments	10	9	7	6	6		

* Judge D disagreed with this group judgment.
** Judge B disagreed with this group judgment.

[5]Judge C was not available to participate in the group discussion and was not involved in the group consensus judgments.


Probability of a Favorable Judgment
Documentation Score with Equal Criterion Weights
Score is percentage of possible rating points awarded.
P
1.0
0.5
0.0
45
50
55
60
65
70
75
80
85
90
Score
ID
Group
Judge A
Judge B
Judge C
Judge D
Judge E