DOCUMENT RESUME

ED 394 953 SP 036 657

AUTHOR Bitner, Ted; Kratzner, Ron
TITLE A Primer on Building Teacher Evaluation Instruments.
PUB DATE Oct 95
NOTE 18p.; Paper presented at the Annual Meeting of the Midwest Educational Research Association (Chicago, IL, October 11-14, 1995).
AVAILABLE FROM Decher Hall, Anderson University, Anderson, IN 46012 ($2).
PUB TYPE Speeches/Conference Papers (150) -- Tests/Evaluation Instruments (160)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Construct Validity; *Content Validity; Elementary Secondary Education; *Evaluation Methods; Evaluation Utilization; Higher Education; Measurement Objectives; Performance Factors; Preservice Teacher Education; *Reliability; *Student Teacher Evaluation; Student Teachers; Student Teacher Supervisors; Student Teaching; *Teacher Effectiveness

ABSTRACT

This paper presents a primer on building a scientifically oriented teacher evaluation instrument. It stresses the importance of accurate measures and accepts the presupposition that scientific approaches provide the most accurate measures of student teacher performance. The paper discusses the scientific concepts of validity and reliability, and provides a checklist for assessing student teacher observation instruments. The checklist suggests two questions relating to validity: (1) Is there ample evidence that the items chosen are related to pupil learning (content validity); and (2) Are the items used in the instrument related to any selected theory of teaching and learning (construct validity)? In reference to the reliability of the instrument, seven criteria are suggested: adequate training of those using the instrument; items stated in overtly behavioral terms; items stated positively; items written in the singular; only one type of behavior per item; items written in present tense; and all items scored using a 5-category forced choice procedure. (Contains 14 references.) (NAV)

A PRIMER ON BUILDING TEACHER EVALUATION INSTRUMENTS

Midwest Educational Research Association

October 11 - 14, 1995

Dr. Ted Bitner
Professor of Education
Anderson University
Anderson, Indiana

Dr. Ron Kratzner
Associate Professor of Education
Anderson University
Anderson, Indiana

A PRIMER ON BUILDING TEACHER EVALUATION INSTRUMENTS

Ted Bitner and Ron Kratzner

1995

Introduction

A recent poll of the public attitude toward public education revealed that parents perceive one of the greatest problems that schools face is the difficulty of getting good teachers. (Kappan, 1993). Teacher preparation faculty members, especially the student teacher supervisors, are often blamed for this difficulty. They are frequently charged with being too lenient in allowing too many ill-prepared people into the field. The supervisor's defense is often that he is simply reporting based on the items on the evaluation instrument in use. The fact is that these evaluation instruments often strongly influence entrance into the teaching profession. The authors' are concerned with what they consider major weaknesses in many of these instruments. [Those entrusted with the creation and use of evaluation instruments for student teachers appear to be either unaware, or deliberating ignoring, tenants of "best practice" in their field. Yet, this area is one that is ignored at great professional peril. The teaching profession is struggling for true professional status. One mark of a professional is to know and use recommended "best practice."

The Quest for Accuracy

The Task of Teachers and Administrators

One of the many tasks supervisors of student teachers have is to make evaluative decisions about the students assigned to them. These supervisory evaluative decisions often have crucial consequences for the student's future. Among the areas affected are: licensure, certification, retention, promotion, and incentive pay. (Houston, 1990)

Because of the importance of these decisions, most student teacher supervisors want to be as accurate as possible. Few supervisors deliberately set out to "get a particular student" or

provide other students with a "free ride." One of the tasks of those in charge of the teacher education program is to see to it that a supervisor's decisions are as accurate as possible. The fact is that the accuracy of the supervisor's decisions often rests, to a great degree, on the accuracy of the evaluation instruments used. This accuracy is often assumed without question. However, once it is questioned, many evaluators may find themselves much less confident than before.

Accuracy and Traditional Scientific Models

There are honest differences of opinion as to how supervisory decisions are best defended. Broadly speaking, these decisions can be classified according to how closely they follow traditional scientific models of evaluation. In most cases, each supervisor must use a designated instrument of unknown pedigree to make final evaluative decisions.

There may be a case which can be made for accurate non-scientific evaluation. And while, at present, the overwhelming assumptions of the education profession knowledge base appears to commit the profession to the canons of scientific research, it does not follow that this should always remain so. However, most administrators are hesitant to make that case. Instead, they wish to argue for scientific "objectivity" when defending a student teacher's evaluation. They usually back up to an assessment made of the student teacher's performance based on what they claim is a relatively "unbiased" evaluation instrument. Seldom does the evaluator defend his evaluation by saying: "Don't question my grade. I simply feel down deep this student teacher is or is not "safe to teach." or "The grade stands because I say so." Once the evaluator volunteers reasons, the assessment of "good reasons" and "bad reasons" is appropriate. The present professional knowledge base criteria identifies "good reasons" with the use of as scientific approaches as possible. It does not seem unreasonable in light of that identification to hold instrument designers as close as possible to "best practice" as defined in the literature.

Importance of Scientifically Oriented Evaluation

Virtually all administrators and supervisors have had at least one preparation course in tests and measurement at some point during their teacher preparation sequence. However, tenants of "best

scientific practice" discussed in this course are often not followed. It may be that many do not feel the principles given in their tests and measurement course were very important. or that the principles discussed extend to evaluation forms in a supervisory context. Some argue since the perfect instrument cannot be developed, we shouldn't try. Granted, there are no perfect evaluation instruments, on the other hand, there are many instruments in use that can be improved enough to provide better information than presently given. It is extremely important for supervisors and administrators to understand how important these principles are and incorporate as many as possible into their instrument design. Teacher evaluation instruments which ignore most of these principles are not only useless for information gathering they are dangerous because they lead to indefensible decisions regarding a student teachers effectiveness precisely at a time when those responsible for decisions are attempting to be perceived as more professional. In our litigious society it is only a matter of time before the basis for these decisions becomes fertile ground for court battles.

Another reason that more scientific principles are ignored is the possibility that many persons responsible for the creation of evaluation instruments simply do not know how to build a scientifically defensible instrument. It is to this latter group that the remainder of this article is addressed for, only those who already grant the importance of scientifically oriented evaluation ask "How can we make the evaluation instruments we use more scientific?"

Criteria for a Scientifically Based Evaluation Instrument

Scientifically inclined designers of student teacher evaluation instruments who are interested in making as accurate as possible statements about students hold two ideas which reinforce accuracy. These two ideas must be clearly understood if scientific based evaluation instruments are to be constructed. These two ideas are **validity** and **reliability**. Because these ideas reinforce accuracy in measurement responsible evaluation instrument design must always consider them.

The most crucial idea in instrument development is validity. *Can we be certain that the items on the instrument evaluate effective teacher' practice?* Validity is like the old joke about the person

whose keys are lost but who looks for them under the lamppost because the light is better there. We may list all kinds of easy-to-test items on an evaluation instrument, but if there is no demonstrated relationship between these items and what we want to find, we are wasting our time. Often the adequacy of our instruments are assumed at face value. Yet, this way of operating is no guarantee of accuracy - scientific or otherwise. In point of fact, the assumption of scientific adequacy cannot be maintained without several important external checks in place.

- Is there ample evidence that the items used are significantly related to pupil learning in the given area being evaluated?
- Are the items chosen related to an acceptable theory of teaching and learning in the given area being evaluated?

The second criteria of responsible measurement is reliability: *Can we be certain that the instrument gives accurate results?* If we cannot measure with accuracy, we are measuring with a rubber yardstick where what we measure one time may vary widely when the evaluation is completed at another time. We can generalize nothing. Among the key questions in establishing reliability are:

- Have those using the instrument been trained adequately enough to interpret the indicators in the same way?
- Are all the instrument's items stated in overt behavioral terms?
- Are all items positively stated?
- Are all items written in singular terms?
- Does each item reflect only one behavior?
- Are all items written in present tense?
- Are all items scored using a 5-option forced choice?

Since the educational scientific research community has identified the ideas of **validity** and **reliability** as the most important checks for accuracy, it may help the instrument designer to have a checklist of several of the more important questions which must be answered about these concepts before accepting any evaluation instrument as valid and reliable enough for use with student teachers. Each of the checks in figure 1 on the following page will be discussed in the remainder of the paper.

figure 1

A CHECKLIST FOR ASSESSING STUDENT TEACHING OBSERVATION INSTRUMENTS

ITEM	DEGREE OF AGREEMENT
A. IS THE INSTRUMENT AS VALID AS POSSIBLE?	
1. IS THERE AMPLE EVIDENCE THAT THE ITEMS USED ARE SIGNIFICANTLY RELATED TO PUPIL LEARNING?	SA MA U MD SD
2. ARE THE ITEMS CHOSEN RELATED TO AN ACCEPTABLE THEORY OF TEACHING AND LEARNING?	SA MA U MD SD
B. IS THE INSTRUMENT AS RELIABLE AS POSSIBLE?	
1.. HAVE THOSE USING THE INSTRUMENT BEEN TRAINED ADEQUATELY?	SA MA U MD SD
2. ARE ALL ITEMS STATED IN OVERTLY BEHAVIORAL TERMS ?	SA MA U MD SD
3. ARE ALL ITEMS POSITIVELY STATED?	SA MA U MD SD
4. ARE ALL ITEMS WRITTEN IN SINGULAR TERMS.	SA MA U MD SD
5. IS ONLY ONE TYPE BEHAVIOR LISTED PER ITEM?	SA MA U MD SD
6. ARE ALL ITEMS WRITTEN IN PRESENT TENSE?	SA MA U MD SD
7. ARE ALL ITEMS SCORED USING A FORCED CHOICE PROCEDURE?	SA MA U MD SD

WHAT IS THE STATISTICAL VALIDITY AND RELIABILITY OF THE INSTRUMENT YOU USE? If unknown indicate this.

ON THE BACK OF THIS PAPER EXPLAIN HOW VALIDITY AND/OR RELIABILITY WAS DETERMINED. If unknown indicate this.

Is the Instrument as Valid as Possible?

There are two types of validity which are usually identified by the designer of evaluation instruments: (1) Content validity and (2) Construct validity. When evidence is gathered which helps the supervisor fixate on items which are most significantly related to increased pupil learning, Content validity is being used. (2) When the instrument is linked to a body of research that is part of an accepted theoretical structure of the discipline, then the instrument has construct validity.

1. Is There Ample Evidence That The Items Chosen Are Related To Pupil Learning? (Content Validity)

Issues relevant to the considerations of content validity include appropriateness of the items, inclusion of enough information to cover the domain of interest, and the level of mastery at which the content is being assessed. Content validity requires that items in the evaluation instrument be representative of the universe of elements in the domain of teacher effectiveness. This domain is the teacher-pupil classroom interaction. Only the quality of this interaction should be evaluated with this type instrument. Other factors thought important should be left to other information gathering instruments. Walberg in Wittrock,(1986) identifies several teacher behaviors positively correlated with pupil learning as well as representative instruments thoughts especially strong. There appears four clusters of effectiveness which must be taken into account. Gliessman (1989) identifies relevant factors of warmth and clarity, as important constructs of learning. in addition, firmness and flexibility are factors which is repeatedly mentioned in the literature. (Kratzner, 1977) Only those items that measure the specified behaviors associated with these clusters should be included in the instrument if the requirement of content validity is to be satisfied. Kratzner and Bitner,(1991) identified the steps in examining the validity of theoretical predictions implicit in an appraisal system.

a) Search research literature for significant studies relating teacher behaviors and pupil outcomes.
b) Organize these studies under each of the 4 clusters mentioned in the literature.

c) Create a list of over 20 possible teacher behaviors which could serve as indicators in each area.
d) Interview students, teachers, administrators, university personnel and create a list of those items ranked highest by all groups
e) Select about 8 of the most frequently chosen behaviors in each area. (This creates a pool of less than 40 items total).

2. Are The Items Used In The Instrument Related To Any Selected Theory Of Teaching And Learning? (Construct validity)

How do I know that the items represent behaviors which are part of an accepted theoretical construct? A review of the literature will help answer this question. If research is being completed and the results are being published, then one can easily determine a body of research that is part of an acceptable theoretical structure. (see Kratzner, 1977)

1. Review existing theories of teaching and learning in the literature.
2. Identify or modify an existing theory of teaching based on the studies reviewed. (This theory should relate various claims in a nomological network in which the teacher behaviors chosen are logical hypotheses.)

It can be seen that validity is a complex concept. Not only must the instrument measure the construct it was designed to measure, it must also be based in a theoretical system which deals with the dimension to be measured. Such instruments are not put together quickly or superficially.

B. Is The Instrument As Reliable As Possible

The second criteria of responsible measurement is reliability: *Can we be certain that the instrument gives accurate results across several uses.* Reliability of behavioral observations are as important as any other type of assessment procedure. (Sattler, 1988) If we cannot identify any assurances of reliability we are measuring with a rubber yardstick. What we measure on one occasion could vary widely on another occasion. One cannot depend on an evaluation of any type if it cannot be expected to produce the same results from one administration to the next (Phillips, 1988) Researchers have

classified possible errors which could occur using a given student teacher evaluation instrument as type I or type II. Type I errors are committed when a student teacher who should not have been allowed to continue is allowed to do so. A type II error is made when the student teacher should have been allowed to continue but is not allowed to do so. When an instrument is unreliable, a type II error is more likely to occur. When this error is present, student teachers who should have been judged favorably tend not to be so judged. This is most damaging to the subject if the results based on the use of the instrument in question is being used for future placement or hiring purposes. According to Gronlund (1986) high reliability is demanded when the decision is important, final, has lasting consequences and where individuals are concerned. Supervisors will find it increasingly difficult to make a case for findings obtained through the use of an instrument with no demonstrated reliability. We can generalize nothing about a particular student teacher using a such an instrument. The importance of questions regarding reliability cannot be overemphasized. Listed below are some of those questions.

1. Have Those Using The Instrument Been Trained Adequately?

This is the question of interrater reliability. One major difference between teacher evaluation and other types of assessment procedure is the importance of establishing observer agreement. The reason for this is that those responsible for accurate assessment must make certain that the assessments given by different observers are a function of the teacher factor and not differences in observer bias. For evaluation to be credible, all the observers must be identifying the same indicators. Again we refer back to the "rubber ruler" analogy. If the observers do not agree when observing the same behaviors, the instrument is being stretched, and the results will not be able to be used reliably. Sattler (1988) lists personal qualities of the observer as a potential source of errors. One can see readily that the personal qualities of the observer, also known as observer bias, can be the cause of a bias. This bias consists of anything the observer does that distorts the assessment of the behavior observed. Some observers may be more limited than others. Different observer' personal theories of teaching may interfere. The tendency to look of some items and ignore other may interfere. In student teaching situations, not only will different observers have different theoretical approaches, but they will have as a basis for a different framework.

The classroom teacher may be looking for a more pragmatic approaches to teaching, while the college supervisor may be looking for behaviors that indicate the student teacher has met the college objectives. The university supervisor from the student's major department may be looking at domain specific teaching behaviors. Obviously, the higher the reliability index, the more confidence we can place on our instrument. With so many different observers, the highest reliability possible is a must if significant meaningful results are to be expected. Several methods for determining reliability are suggested (Rosenshine and Furst, 1973)

One of the most common methods for determining interrater reliability is listed below:

1. Use the split-half method (Roscoe, 1975). In this approach, divide the instrument into two parts, usually by separating the odd numbered items from the even ones.
2. Compute a correlation between the two parts.
3. Reordered using the two halves like two tests. A Pearson correlation coefficient is computed between the two means of the halves.

2. Are All Items Stated in Overtly Behavioral Terms?

Medley and Mitzel (in Gage, 1963) suggested that sign system tend to predict gain better than other systems. They explain a sign system as one that lists beforehand a number of *specific acts of behavior* which may or may not occur during a period of observation. Rosenshine (1971) argues that "observation systems can...be classified according to the amount of inference needed by the observer or the person reading the research report.." .The term "inference" refers to the process of interpretation needed between the objective behavior seen or heard and the coding of this behavior on an observational instrument. Low inference measures focus on specific, relatively objective behaviors such as 'teacher repeats student ideas', or 'teacher asks evaluative questions'. Gellert (1955) explained, "the less inference required in making classifications, the greater will be the reliability (, p. 184). Since the highest possible reliability is desired, low inference items should be used. The lowest inference items tend to be a record of the overt behaviors of the person being observed.

An example of low inference items used to measure a student teachers writing might be 'writing on board so that it can be read easily." This requires a much lower inference than an item such as "uses professional language both orally and in writing. In the former example, one knows when writing is legible or it is not, while the latter says nothing about legibility per say. Although most supervisors might believe that use of professional language is certainly desirable what constitutes "professional" language demands a high inference. Is legibility a factor in "professional" writing or not?

1. Write the item in its rough draft form.
2. Ask What overt behavior can I point to which allows me to evaluate the presence of this concern?
3. Rewrite the item in the most overt behavioral terms. possible.

3. Is Each Item Stated Positively

Medley and Mitzel (in Gage, ed.). report that "teacher ignores pupil contribution" is a better item than "teacher fails to recognize significant contributions." "Displays knowledge of subject matter is not as strong as "avoids mistakes in knowledge." Some of the ways to state items more positively include the following:

1. Avoid the use of terms such as "does not," "failed to" "can not."
2. Use identifiers such as promotes, establishes, etc.

4. All Items Written In Singular Terms.

Medley and Mitzel (in Gage, 1963) point out that better items are expressed in singular terms. We observed the teacher using a specific game as a practice activity. Thus, the item is "Using game for practice (singular) rather than "Using games for practice" (Plural) If several games were played in a session, the item is keyed as strongly agree rather than mildly agree.

1. Check each item for singularity
2. Change plural items to singular ones.

5. Is only one type of behavior included per item?.

Writing an item that contains several behaviors confuses both the evaluator and a person who subsequently reviews the evaluation. Which of the several items was observed and noted? Were they all? If the evaluator believes that making eye contact

with students demonstrates positive interaction. The item should be: "Makes eye contact with pupil" If another important behavior is calling pupil by name, a separate item should be "calls student by name." Contrast the above items with the following from an evaluation instrument for student teaching currently in use: "demonstrates positive interactions by smiling, looking at students, calling student by name, helping students with problems, and complimenting students" -all in the same item.

1. Identify the specific behaviors in the item.
2. Separate each overt behavior into a new item.

6. Are All Items Written In Present Tense?

Using present tense in items such as "Looks at pupil often, makes use of student idea" allows the evaluator to be in sync with the classroom environment.

1. Identify behaviors
2. Make a list of the present tense verbs such as make, use, identify.

7. Are All Items Scored Using a 5 Category Forced Choice Procedure?

If two few categories are used, fine distinctions will be lost (Sattler, 1988) A rating scale that simply asks for the positive or negative presence of a behavior other than "not observed" will give only dichotomous results at best. Often it is more helpful to identify the degree a behavior is present. The fact that a student teacher praised or did not praise is often less important than the quality of the praise. Another problem with too few categories is that there is a tendency for the supervisor to mark the central scoring categories more than the end scoring categories. On the other hand another source of error contributed to by the instrument is the fact that there may be too many categories in the system that must be scored on a given occasion. Page (1974) found that a 5-option forced choice scale was quite resistant to the attempts to produce bias.

1. Adopt a Likert-type scale.
2. Use an odd number of categories
   (5 recommended).
3. Write the scale above each category to be scored.
   (Strongly agree; mildly agree; undecided;
   mildly disagree; strongly disagree)

Summary

This article has presented a primer on building a scientifically oriented teacher evaluation instrument. It established the importance of accurate measures, the presupposition that scientific approaches provide the most accurate measures, discussed the scientific concepts of validity and reliability and provided ways to increase these concepts as much as possible in an student teacher evaluation instrument. It is designed to help those designers of evaluation instruments who desires a more scientifically defensible instrument but are unclear as to how to approach the task.

WORKS CITED

Andrews, T.E. & Barnes, S. (1990) "Assessment of teaching" in Houston (ed.) **Handbook on Research in Teacher Education.** New York: Macmillan. p. 573

Gellert, Elizabeth.(1955) "Systematic Observation: A Method in Child Study." **Harvard educational review** 25: 179-195.

Gliessman, David H. et al. (1989)"Applying a Research-Based Model to Teacher Skill Training." **The journal of educational research.** 83 (November/December, 1989),69.

Gronlund, Norman E.(1986) Measurement and evaluation in teaching, 5th ed. New York: Macmillan. p.87

Kratzner, Roland R.(1977) **A teacher behavior instrument for a multidimensional teaching model.** unpublished doctoral dissertation, Muncie, Indiana.

Kratzner, Roland R. and Ted Bitner (1991) "Validity and Reliability Measures for a Student Teacher Evaluation Instrument " **The teacher educator.** 27 (Summer 1991): 38-43.

Medley, Donald M., and Mitzel, Harold E. "Measuring the Classroom by Systematic observation."(1963). In **Handbook of research on teaching,** Edited by N. L. Gage. Chicago: Rand, McNally and Company, Inc. 247-328.

Phillips, J. L.., Jr. (1988). **How to think about statistics.** New York: W. H. Freeman and Co., 44-66.

Page, Collin.(1974) **Student evaluation in teaching: the american experience.** London: Society for Research into Higher Education

"The 25th annual Phi Delta Kappa/Gallup poll of the public's attitudes toward the public schools". (1993) Phi Delta Kappan. Oct. 1993, 75, 5. 137-152.

Roscoe, S. T.. (1975) **Fundamental Research statistics for the Behavioral sciences.** 2nd. ed. New York: Holt-Rinehart-Winston. p.173

Rosenshine, Barak, and Furst, Norma.(1973) "The Use of Direct Observation to Study Teaching." In **second handbook of research on teaching**. Edited by Robert M.W. Travers. Chicago: Rand, McNally College Publishing Company. 169, 273.

Sattler, J. **Assessment of children.** (1988) 3rd ed.. San Diego: Sattler. 25-31.

Walberg, H.(1986) "Synthesis of Research on Teaching". in Wittrock, Merlin C. (ed.) **Handbook of research on teaching**, New York: Macmillan Publishing Company.

THE KRATZNER-BITNER TEACHER BEHAVIOR DIAGNOSTIC INSTRUMENT

Teacher _____ Observer _____

Date _____ School _____ Time _____ Subject _____

DIRECTIONS: Please indicate the degree to which you agree with the statement below

SA - strongly agree
MA - mildly agree
U - undecided
SD - strongly disagree
MD - mildly disagree

I. AFFECTIVE CONCERNS

A. The tendency to show concern for pupils feelings (warmth) indicated by:

1. _____ avoiding talking to pupil like a "little child"
2. _____ smiling often
3. _____ avoiding doing or saying things which may hurt pupil
4. _____ looking at pupil often
5. _____ using pupil name often
6. _____ making use of pupil ideas
7. _____ saying nice things to pupil
8. _____ discussing problems pupil is having

COMMENTS:

B. The tendency to provide firm, supportive feedback to pupils (firmness) indicated by:

1. _____ saying what to do in order to avoid making mistakes again
2. _____ telling pupil when a mistake is made.
3. _____ avoiding using "gutter" language
4. _____ following through on any promises or threats made
5. _____ avoiding using sarcastic tone or remarks or shaming pupil
6. _____ apologizing when teacher has made a mistake
7. _____ avoiding scolding pupils who have not done wrong
8. _____ avoiding slapping or hitting pupil

COMMENTS:

II. COGNITIVE CONCERNS

A. The tendency to communicate directly what is expected of pupils (clarity) indicated by:

1. _____ speaking so all words can be heard easily
2. _____ avoiding mistakes in information given
3. _____ reminding pupil of important ideas to remember in a few sentences
4. _____ telling pupil what to expect in behaviors he is to use
5. _____ emphasizing important words spoken
6. _____ relating new information to information already given
7. _____ avoiding talking too fast or too slow
8. _____ letting pupil know when he has done something correct by repeating it

COMMENTS:

B. The tendency to elicit a variety of responses from pupils (flexibility) indicated by:

1. _____ using AV aids (movies, slides charts, graphs, pictures)
2. _____ writing on blackboard so it can be read easily
3. _____ asking a variety of questions at different levels of difficulty
4. _____ using games for practice activity
5. _____ Materials and activities appropriate to pupil interest and level
6. _____ activities short enough to keep attention
7. _____ High pupil involvement in analysis or creative activities
8. _____ Some value-clarification techniques used

COMMENTS:

COMMENTS

WARMTH

FIRMNESS

CLARITY

FLEXIBILITY