KEY POINTS IN
EDUCATIONAL
EVALUATION

Hartono

This short article will examine the key points in educational evaluation. It will begin with the discussion about the concept of evaluation, the areas of evaluation, the general principle of evaluation, the basic ways of evaluation procedures, the characteristics of a good instrument of evaluation, the planning of a testing program, the test administration, the interpretation of test scores, item analysis, the report of evaluation, and, finally, conclusion will end this article.
What is evaluation?

Many authors have attempted to clarify the concept of evaluation as opposed to measurement. This happens because in some instances evaluation is used as a synonym for the term measurement (Gronlund, 1976), and the two terms are often used with little regard for their meanings (Remmers, et al., 1960).

In what respects are the two terms different? Hopkins and Stanley (1981) state that evaluation refers to a summing-up process in which value judgement play a large part, as grading and promoting students, whereas the development, administration, and scoring of the test constitutes the measurement process. Interpreting such scores is, then, part of the process of evaluation. In attempting to distinguish the terms, Gronlund (1976) pointed out that evaluation includes both quantitative and qualitative descriptions of pupils plus value judgement concerning the desirability of that behaviors. Measurement is, on the other hand, limited to quantitative descriptions of pupil behaviors. Thus, evaluation is much more comprehensive and inclusive term than measurement. Remmers et al. (1960) differentiated the two terms in somewhat similar way. Measurement, according to them, refers to observations that can be expressed quantitatively and answers the question "how much"; whereas evaluation goes beyond the question of how much, to concern itself with the question "what value".

What is to be evaluated?

In general, the areas in which evaluation of pupil growth and development may be made can be grouped into intelligence, interest, achievement, physical, and emotional (Remmers, et al., 1960). According to Wandt and Brown (1957) evaluation is comprehensive because evidence is obtained regarding pupils' abilities, interest, health, adjustment, achievement character - in fact, every aspect of the personality.

Furthermore, Remmers et al. (1960) described nine areas within which the teacher needs Information for adequate pupil evaluation, namely scholastic aptitude, scholastic achievement, special abilities, personal interest and plans, health physical status, home and family relationships, emotional and social adjustment, attitude, and work experience.

What are the general principles of evaluation?

According to Gronlund (1976) there are at least five general principles of evaluation, namely (i) Determining and clarifying what is to be evaluated always has priority in the evaluation process; (ii) Evaluation techniques should be selected in terms of the purposes to be served; (iv) Proper use of evaluation techniques requires an awareness of their limitation as well as their strengths; and (v) Evaluation is a means to and end, not an end itself.

What are the basic ways of evaluation procedures

Gronlund (1976:20) presents a table that summarizes the basic ways of describing classroom evaluation procedures as follows:

<table>
<thead>
<tr>
<th>Basis for Classification</th>
<th>Types of Evaluation</th>
<th>Function of the Evaluation</th>
<th>Illustrative Instruments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nature of Measurement</td>
<td>Maximum performance</td>
<td>Determine what a person can do when performing at his best.</td>
<td>Aptitude tests; Achievement tests;</td>
</tr>
<tr>
<td></td>
<td>Typical performance</td>
<td>Determine what a person will do under natural condition.</td>
<td>Attitude scales; Interest inventories; Observational techniques; peer appraisal</td>
</tr>
<tr>
<td></td>
<td>Placement</td>
<td>Determine what possession of prerequisite skills, degree of mastery skills of course objectives, and/or best mode of learning.</td>
<td>Readiness tests; Aptitude tests; Pretest on course objectives Self-report inventory; Observational techniques</td>
</tr>
</tbody>
</table>

52

BESTARI, MEI·JULI 1993
**What are the characteristics of a good instrument of evaluation?**

Unlike the previous items, this item 5 will be discussed in more detail. Remmers et al. (1960) has pointed out that an evaluation instrument is judged for its adequacy, efficiency, efficiency, and consistency as a measuring device on the basis of commonly accepted qualities. These qualities are validity, reliability, administerability, and interpretability.

**Validity**

Even though many authors have proposed the concept of validity, the basic idea is basically just the same, that it is to say the validity of an evaluation device is the degree to which it measures what is intended to measure. There are several types of validity, and many textbooks discuss them in different ways. Remmers et al., for instance, proposed four types of validity, namely content validity, concurrent validity, and construct validity.

- **Content validity.** Content validity is evaluated by showing how well the content of the test samples the subject matter about which conclusions are to be drawn. It is also known as "face validity" and "logical validity" and is described by the relevance of a test to different types of criteria.

- **Concurrent validity.** Concurrent validity is evaluated by showing how well test scores correspond to already accepted measures of performance made at the same time.

- **Predictive validity.** Predictive validity is evaluated by showing how well predictions made from the test are confirmed by evidence gathered at some sequence time. It is quite similar to concurrent validity, except that the evidence on the criterion measured used is collected later.

- **Construct validity.** Construct validity is evaluated by investigating what psychological qualities a test measures, or in other words, by demonstrating that certain explanatory constructs account for performance on the test.

There are some factors that affect test validity. The evaluators should recognize the factors that test to make tests valid for their purposes. Among them are (i) cultural influences, (ii) response sets, (iii) increased relabeled which causes the loss of validity, and (iv) difficulty of items in directions to pupils.
Reliability

Reliability is the consistency with which a test yields the same results in measuring whatever it does measure. There are some methods of estimating reliability, all of which involve some ways of obtaining at least two measure with the same instrument and determining the agreement between them. They are as follows (Gronlund, 1976:122):

Test-retest method. Typical provides medium to large reliability coefficients for a given test, may be larger split-half method if time interval is short. Coefficients become smaller as time interval between tests is increased.

Equivalent-forms method. Typically provides medium to large reliability coefficients for a given test. Tend to be lower than test-retest method using short time interval. Or, if the method is with time interval, it typically provides smallest reliability coefficients for a given test. Coefficients become smaller as time interval between tests is increased.

Split-half method. Typically provides largest reliability coefficients for a given test. Spuriously high estimates are produced for speed tests.

Kuder-Richardson method. Typically provides reliability estimates are smaller than those obtained by split-half method. These estimates are also inflated by speed.

Administerability

The administrability of evaluation device refers to the ease and accuracy with which the directions to pupils and evaluator can be followed.

Interpretability

The interpretability of evaluation device refers to how readily scores may be derived and understood.

How should a testing program be planned?

Remmers et al. (1960) suggested activities that can contribute to effectiveness in planning a new testing program or revising a program already in operation. They are (i) secure cooperation from the entire staff; (ii) Determine the purposes of the program so it will fit the various needs of the specific school; (iii) Find out about available instruments; (iv) Decide what kind of instruments will best fulfill the purpose and either select them from available tests or develop them within the school; (v) Make plans for administering the testing program and set up a tentative schedule for the year; (vi) Train the personal necessary for the administration, scoring, and analysis of test results; and (vii) Put the program into operation and use the result to achieve the purposes.

How should tests be administered?

In a smaller testing program, administering a test or an evaluation device involves several steps. Among them are (i) avoiding distractions, (ii) distributing tests to examiners, (iii) adhering to directions, (iv) timing, (v) supervision, (vi) ending the test, (vii) making notation, (viii) security of examination.

How should test scores be interpreted?

As has been done in Item 5, this item will also be discussed in sufficient detail. The simple methods of interpreting test scores include the following:

Simple ranking. One somewhat easy and useful method is that of ranking. Suppose Mary got 70 on English test. What does this mean? If the evaluator knows the score of the other members of the class he can place her according to her relative position. This may be done by arranging the scores in order, from highest to lowest, and assigning each score a number according to its position. Usually he gives the highest score a rank of one, the second highest a rank of two, etc.

<table>
<thead>
<tr>
<th>Pupil</th>
<th>J</th>
<th>F</th>
<th>W</th>
<th>L</th>
<th>M</th>
<th>I</th>
<th>S</th>
<th>E</th>
<th>G</th>
<th>U</th>
<th>K</th>
<th>A</th>
<th>N</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score</td>
<td>80</td>
<td>78</td>
<td>75</td>
<td>74</td>
<td>70</td>
<td>68</td>
<td>65</td>
<td>62</td>
<td>60</td>
<td>58</td>
<td>55</td>
<td>50</td>
<td>48</td>
<td>46</td>
</tr>
</tbody>
</table>

Test score reading

When the scores are arranged in order, the following arrays appears:

<table>
<thead>
<tr>
<th>Pupil</th>
<th>J</th>
<th>F</th>
<th>W</th>
<th>L</th>
<th>M</th>
<th>I</th>
<th>S</th>
<th>E</th>
<th>G</th>
<th>U</th>
<th>K</th>
<th>A</th>
<th>N</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
</tr>
</tbody>
</table>
Be sure that where two or more scores are alike, the places they would otherwise hold should be given as rank to those with the same scores.

**Percentile ranking.** The most essential disadvantage of the method of ranking is that the procedure takes no account of differences in the size of the group. For example, a rank of 10 in a group of 10 means quite a different thing than a rank of 10 in a group of 100. In the first case the rank is the lowest in the group, whereas in the second it is one of the best. And one method that eliminates this difficulty is that of percentile ranks.

Percentile ranks differ from simple ranks in that they express the position of any score in the group in terms of the percentage of the group below that score. In the earlier example it was stated that Mary's score was 70 gave her a rank of 70 or a rank of 5 places her above 10 others. Putting this in another way, it can be said 10/15 or 66.67% of the class makes a lower score than Mary on the test.

**Central tendency or averages.** Among the kinds of averages or central tendency most often used in educational work are the arithmetic mean and the median (Noll, 1957). The arithmetic mean is obtained by adding all the scores and dividing by the number of scores. The median is simply the middle score above it, and equal number of score half and the upper half of the group and the highest score of the lower half.

**Semi-interquartile range.** It is a measure of the spread of the middle half of the scores in any distribution, with the formula:

*Standard deviation.* It is a measure which is based on the deviations of scores, and it is the most reliable because it takes into account the actual variation of each score from the mean.

There are two concepts of measurement related to the interpretation of test scores. They are norm-reference measurement-comparing a pupils' performance on a test to criteria that were established in the structural objectives (Olivia, 1982).

**Why item analysis?**

Anastasi (1982) points out that knowing the basic concepts and techniques of item analysis can help test users in their evaluation of published tests. In addition, item analysis also relevant to the construction of tests prepared by teachers for classroom use. Effective item writing can materially improve classroom tests.

**Why reporting to pupils and parents?**

The main reason for reporting to pupils and parents is to facilitate the learning and development of pupils. Thus, reports should (i) clarify The objectives of the school program, (ii) indicate the pupil's strengths and weaknesses in learning, (iii) promote more understanding of his personal-social development, and (iv) contribute to his motivation (Gronlund, 1976:512).

**Concluding remarks**

This article cover a wide range of discussion about the essential point in educational evaluation. However, with the elimination of the pages only two items (5 and 8) are treated somewhat in detail.

**References**


