《教育行政與評鑑學刊》
2006年12月，第二期，頁43-58

教師表現評鑑的信效度議題

黃財尉

摘　要

信度與效度是精確評鑑不可或缺的考量要項，特別是針對教師表現的評鑑，本文探討許多評鑑過程中常見的誤差，也介紹許多被視爲能降低誤差有效的解決方案，經由對評鑑誤差知識的增加，更能期待增進評鑑者的評鑑技巧並避免可能的評鑑誤差。

關鍵詞：信度、效度、教師表現評鑑

黃財尉：國立嘉義大學輔導與諮商學系副教授
電子郵件　twhuang@mail.ncyu.edu.tw
收件日期　2006.10.9；修改日期　2006.10.24；接受日期：2006.11.23

Journal of Educational Administration and Evaluation
December, 2006, Vol. 2 pp.43-58

Validity and Reliability Issues for Teacher Performance Evaluation

Tsai-Wei Huang

Abstract

Reliability and validity should be taken into account in the matter of the whole consideration of evaluation accuracy, especially in the evaluation of teacher performance. This article discussed some evaluation errors and introduced some solutions that were found effectively to minimize certain errors. Through increasing knowledge of appraisal theories as well as possible errors during the rating processes, appraisers are more likely expected to improve their appraising abilities, identify sources of errors, and avoid many possible rating errors.

Keywords: reliability, validity, teacher performance evaluation

Tsai-Wei Huang：Associate Professor of Department of Counseling, National Chiayi University
E-mail：twhuang@mail.ncyu.edu.tw

Manuscript received: 2006.10.9; Revised: 2006.10.24; Accepted: 2006.11.23

Introduction

Assessing the classroom performance of teachers, just like evaluating personnel performance in an organization, is a prerequisite for assuring the quality of teaching. It plays a very important role in ensuring the proper management of a school's human resources to increase teachers' productivity. Through the quality control of classroom performance assessment for teachers, it is much more expected that a proper teacher will be selected by a needed school.

However, a wide variety of factors, such as raters, instruments, tasks, etc, would affect the quality of the evaluation of performance (Austin & Villanova, 1992). To overcome appraisal errors and achieve the quality purpose effectively, the appraisal instrument should make performance measurement both reliable and valid. Indeed, accuracy and consistency of evaluation are the most critical premises during the process of judgment, and the issues of validity and reliability in assessing the teacher classroom performance are related to the increase of teachers' productivity.

Reliability is regarded as a premise of validity. A reliable data set is not necessarily a valid one, while a valid measure should yield reliable data about what it is concerned with. For instance, consider a manager who always rates his subordinates' performances in a bad mood. The result is consistent, in this case, but the appraisals can never be regarded as valid. Another distinction between reliability and validity argued by Latham and Wexley (1994) is that reliability is an attribute of one factor of rating, e.g., a job performance rating, while validity is the relationship between two factors of rating, e.g., how a performance rating correlates with another independent measurement of performance. In other words, reliability is necessary, but not sufficient for validity. The former, however, is an essential prerequisite for the latter.

Reliability

Reliability of appraisal can be considered through three facets of consistency:

teachers, raters, and criteria. The first, consistency in the performances of teachers, also called stability of performance, is assessed through the test-retest method to compare one time period to another. In teacher evaluation, this requires measuring the performance of the same teacher on two or more occasions with the same performance appraisal instrument. Practically, teachers may be evaluated in a regular-session occasion and in an emulating occasion where the same unit is instructed to the same audience and the same performance criteria is inspected. Correlation between appraisals from different periods of time is used as an index of reliability. The higher the degree of similarity from one time the other, the higher the reliability of the evaluation of performance. Due to inevitable fatigue or teaching fluctuation, an ideal reliability in this estimating method is expected to be 0.70 or higher.

The second dimension of consistency emphasizes consistency among the raters' appraisals. The corresponding reliability for this kind of consistency is called inter-rater reliability. It is preferable that two or more raters should be in high agreement when evaluating a teacher independently. Linn and Gronlund (2000) suggested that Generalizability theory can be used to evaluate consistency across raters or across tasks, where G and Phi coefficients can provide the stability of related absolute decisions, respectively.

The third method assessing reliability focuses on the internal consistency of the items that comprise a scale. It answers the question of whether all the items on a scale are assessing the same dimension or quality (Latham & Wexley, 1994). The value of consistency can be calculated through split-half method or Cronbach alpha method, that is, by calculating the correlation between odd- and even-numbered items on a rating scale or calculating the value of Cronbach alpha between each item and the total score. A preferred value of reliability is expected to be more than 0.80.

Validity

The validity of a measure can be regarded as how well it fulfills the function for

which it is being used. That is, it is meaningful only in terms of its specific uses. As Hopkins (1998) argued, validity is related to correct inferences made from performance in the measure. These inferences pertain to performance on a universe of items (content validity), performance on some criterion (criterion-related validity), and the degree to which certain psychological traits are actually represented by performance (construct validity).

An appraisal instrument is thought to have content validity if it contains a representative sample of the teachers' performances. That is, this appraisal instrument can reflect what it intends to evaluate about teachers' performances. The extent of content validity is qualitative-oriented and is judged by experts after job analysis. Namely, a job analysis shows whether the appraisal instrument has content validity with revealing the extent to which the teacher is evaluated only on job-related factors. If unrelated job factors are evaluated, unnecessary legal issues may be happened. In a classroom situation, for example, the extent of how accurately to evaluate a teacher's teaching performance might fall into four job domains: knowledge of subject contents, ability of teaching and learning psychology, skills of creating learning environment, and individual professionalism (Dwyer & Villegas, 1993). Evaluations beyond these might arouse legal controversies.

In contrast to content validity, criterion-related validity, especially predictive validity, is purely an empirical matter. Predictive validity is used to predict teachers' future performance on a different job. However, as Latham and & Wexley argued (1994), it is seldom used in organizations because the validation sample of teachers is required more than thirty people.

Construct validity is evaluated by investigating what qualities an appraisal of performance measures. The qualities are psychological traits or abilities that are unobservable, yet exist in theory. A multi-trait-multi-rater framework, similar to the multi-trait-multi-method approach (MTMM, Campbell & Fiske, 1959), is usually used to assess the construct validity of appraisal decisions. High correlations within a

criterion/construct among different raters are expected to reach convergent validity, while lower correlations between different criteria/constructs are expected to ascertain discriminant validity.

It should be noticed that the categories of content validity, predictive validity, and construct validity are merely for the sake of convenience. In fact, in an interpretative perspective, categories of validity should not be viewed as separate concepts; rather, they should be considered in a unitary concept (Messick, 1989, 1995). That is, validity can be an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and representativeness of content; rating rubrics match internal structure, results generalize across groups and tasks; the criterion relevance is evidenced; and finally value implications of evaluation based on test scores or other modes of assessment is appropriately interpreted. Thus, even though validity is classified into several categories, when we interpret the appraisal scores, it is still necessary to have a comprehensive view and account for the content, criterion-related, and construct validity. Furthermore, to concentrate on appraisal validity only and to avoid unnecessary debates, managers still need to aware some appraising principles, such as the due process, privacy, equality, openness of procedures to public, humanness, client benefit, academic freedom, and respect for autonomy (Peterson, 1995).

Problems in Appraisal

With desirable reliability and validity, an appraisal can be viewed as valid and accurate. In practice, however, the objective of accuracy for an appraisal is not easy to reach. Attributing the errors of appraising merely to the rater facet is not a comprehensive view. In fact, factors in the work environment can interfere with the appraiser's evaluation of the teacher's performance (Hedge & Laue, 1988). Several factors or nuisances always diminish the degree of the accuracy of an appraisal. Some factors that influence the level of accuracy can be seen from the analysis of reliability. For instance, the stability of an appraisee's performance, the consistency of appraisers, and the quality of the instrument as well as the dimensions of evaluated traits would

affect the test-retest reliability, the inter-rater reliability, and internal consistency, respectively. Besides that, some factors from the analysis of validity can potentially influence the accuracy of an appraisal; for example, the purpose of evaluation, the criterion-related measures, and the interpretation as well as the use of appraisals will always determine the level of accuracy of an appraisal. To sum up, the errors of appraisal may come from the aberrances in (1) the purpose of an appraisal, (2) the teacher's performance, (3) the instrumentation, (4) the evaluated dimensions or traits, and, most importantly, the rater factor, including (5) the interrater and (6) the intrarater factors. Namely, the first three errors refer to the manager factor and the rest refer to the rater factor.

1. Purpose of appraisal

For any performance appraisal, the main purposes have to be determined. They might be related to an teacher's retention, termination, promotion, demotion, transfer, salary increase or decrease, or admission into a training program (Latham & Wexley, 1994). Different purposes of appraisal would affect an appraiser's evaluation of an teacher's performance on two appraisal situations, even though the teacher's performance is identical during the evaluation processes. It is reasonable, for instance, that an appraiser would give an appraisee a more lenient evaluation for the purpose of training than that for the purpose of termination. The purpose-dependent appraisal, indeed, influences the accuracy of evaluation.

2. Teacher's performance

Appraisals cannot be implemented in such a way as to be completely consonant with an teacher's performance; instead, they are regularly held after a certain period of time. All are evaluated, therefore, on the basis of fragile performances that occur at the critical time of appraisal. Some teachers perform more poorly during that appraisal period than usual because of, for instance, fatigue or anxiety, whereas others will do a better job than usual on account of the Hawthorne effect. Appraisals would be affected by the appraisees' unstable performance, and hence lose some of the accuracy of evaluation.

3. Instrumentation

The instrument or available tools and the appraisal are closely related to each other. Yet, the selection of appraisal instrument is often governed by expediency rather than relevancy, especially under the common limitations of time and money (Weitz, 1961). Appraisers may borrow or use; borrow, retrofit, and use; or hastily develop and use an instrument to appraise teachers' performances. It is not logical to assume that the appraiser inherits a tool adequate for his/her needs (Hedge & Laue, 1988). In this case possible inaccurate ratings are due to the inadequate instrument, not to the appraiser. Thus, the appraiser will evaluate an teacher's performance inaccurately through the selection of an inappropriate instrument.

4. Dimensional or Trait Differences

Since some dimensions display concrete characteristics that provide more tangible observable behaviors, they appear inherently easier to rate than others. Thus, differential levels of accuracy might be obtained for different dimensions when rating an appraisee (Landy & Farr, 1975; Hedge & Laue, 1988). This point of view has been demonstrated by Cronbach and Gleser (1957), who showed that broad and global constructs allow the appraiser to predict a variety of behaviors at moderate levels of accuracy, whereas narrow and specific constructs allow the appraiser to evaluate at high accuracy levels within a limited range of behaviors but do quite poorly outside that range. For instance, to appraise the dimension of a teacher's attitude, the statement "attitude display carelessly in teaching" is more ambiguous than the statement "this teacher could not be expected to pay attention to students' responses when someone rose his/her hand." The latter generates a concrete behavioral incident and may create high levels of appraisal accuracy, while the former necessitates much developmental time and energy to identify the necessary anchors and, thus, may be found to possess much lower levels of appraisal accuracy. Therefore, as Hedge and Laue (1988) argued, if dimensions differ along a continuum from the abstract to the concrete, rating accuracy can be affected, and blame should not arbitrarily be placed on the appraiser.

5. Individual Differences Between Appraisers

An appraiser's characteristics are related to his/her ability to judge others. Taft (1955) identified factors of age, high intelligence, good emotional adjustment, and social skill, which are consistently correlated with the ability to judge the personality characteristics of others. Borman (1979) found that the most consistently high correlates of accuracy were intelligence, personal adjustment, and detail orientation. In a Dickson, Hedge, and Teachout (1987) study, general ability and situational constraints were found to be significant predictors of rating accuracy. Based on these evidences, rating accuracy might be viewed as related to appraiser individual differences. Therefore, appraisers' characteristics might result in discrepancies in the appraisal of the performances of the same appraisees.

6. Rating Error Within an Appraiser

In the context of the workplace, rating errors are defined as the influence of factors other than the teacher's job performance that systematically change the appraisal of that teacher. In this case, rating errors would reduce the appraisal validity. There might be ten rating errors frequently occurring in teacher performance process (Latham & Wexley, 1994; Hopkins, 1998):

A. Contrast-Effect Error

The contrast-effect error is the tendency for a rater to appraise a person relative to other individuals rather than to the requirement of job. Thus, an average-performance teacher may be rated showing as lower performance when compared with other high-performance teachers.

B. First-Impression Error

First-impression error occurs when a manager makes an initial favorable or unfavorable judgment about an teacher and then ignores subsequent information so as to support the initial impression. That is, raters acquire an impression of the teacher's performance based on the initial criterion that colors their judgment of subsequent performance.

C. Halo Effect

The halo effect is the tendency, when judging a person's perfor mance on the job, to be influenced by other aspects of the person's job performance or by a general impression of the person. For example, a person who is excellent in only one area of the job may be rated inaccurately as outstanding on all areas of the job.

D. Order Effect

In one study, appraisees that were evaluated earlier tended to receive higher ratings than those appraised near the end of the sequence. This may be due to weariness in the appraiser's physical and mental condition.

E. Similar-To-Me Error

Raters tend to judge more favorably those persons whom they perceive as similar to themselves. That is, the more closely an teacher resembles the rater in values, predisposition, or background, the more likely the rater is to judge that individual favorably. By contrast, the *dissimilar-to-me error* refers to the strong tendency to give the appraisee unlike the appraiser a low evaluation. These judgments are not related to job performance.

F. Central-tendency Error

Appraisers want to play it safe and therefore consistently rate an teacher on or close to the midpoint of an appraisal scale when the teacher's performance clearl warrants a substantially higher or lower rating.

G. Negative and Positive Leniency Errors

This error refers to a manager who is either too severe or too general in ratin teachers. This would cause bias in regard to an teacher's performance.

H. Logical Error

This error refers to an appraiser who illogically infers that one aspect of teacher's performance related to other aspects of performance. For example, if

manager conceives communication ability as highly related to selling goods, he/she may evaluate the teacher as a good seller based on a high rating on his/her communication ability.

I. Response Set

This error refers to an appraiser who, when encountering an uncertain situation, tends to judge an appraisee's performance significantly severely, leniently, or moderately. For example, when a manager is unaware of his/her teacher's performance on a specific job, he/she tends to rate the teacher close to the midpoint of the appraisal scale.

Possible Solutions

How to solve the problems described above, to minimize the damage of errors that limit appraisal accuracy is the most important question that a manager and an appraiser faces. Regarding to the errors coming from the manager factor, a manager need to clarify what goals he/she intends to achieve and should avoid putting different purposes in an appraisal. Telling teachers the specific objectives, motivating them, and regularizing appraisal as usual would help managers free from some administration mistakes. Of course, improving the knowledge of instrument, keeping contact with publishers, and attending some workshops of instrument would help managers hit the target.

Regarding to the errors coming from an appraiser, psychologists have stressed the importance of providing training to improve objectivity and accuracy in evaluating an appraisee's performance, of which some cognitive schemata that interfere with an appraiser's ability of making an accurate or valid appraisal can be minimized. The contents of training programs for an appraisee may include a lecture, group discussion, active participation, knowledge of evaluation results or feedback, and being given a chance to practice by observing and rating videotaped individuals (Latham & Wexley, 1994). Bernardin and Buckley even concluded correctly that an effective rater training

program should concentrate on enhancing the accuracy of rating through discussion of the multidimensionality of work performance, the importance of recording objectively what is seen, and the development of specific examples of effective and ineffective teachers (see Latham & Wexley, 1994, p.155). To achieve the requirement of effectiveness of rater training, Latham, Wexley, & Pursell (1975) suggested that evaluation training might provide exercises for trainees by showing videotapes of job candidates being evaluated and asking them to rate both the manager and candidates shown in the videotapes. These processes would help appraisers focus their attentions on accurate statements of traits and improve the consistency of ratings.

Finally, to minimize rating errors within an appraiser, some exercises were suggested by Latham and Wexley (1994, pp. 156-158). These errors refer to the similar-to-me error, the halo error, the contrast error, the first impression error, and the positive and negative leniency error.

The first exercise focuses on the similar-to-me error. Among the many possible solutions brainstormed by the trainees for minimizing error in performance appraisals are: (1) Establish standards of performance expected on all jobs before rating teachers; (2) Make certain that all criteria on which teachers are evaluated are clearly job related; (3) Rate teachers solely in relation to their job responsibilities, not in terms of how similar they are to oneself; (4) Have teachers evaluated by multiple raters with different backgrounds and attitudes.

Second, possible suggestions provided to solve the halo error are: (1) Do not listen to comments about a person until your own evaluation has been completed; (2) When an individual is to be evaluated by multiple raters, be certain that the raters assign their ratings independently; group discussion about the teacher should come after everyone has had an opportunity to observe and evaluate the individual; (3) Rate the individual solely on the behavioral items that define a given criterion. Recognize that different performance measures are not always related. A person can do well on one criterion and perform poorly on another (e.g., a professor may be a good researcher and a poor

teacher).

Third, for the contrast error possible solutions to are: (1) Appraise a large number of people at the same time; the error is more frequent when only a few individuals are interviewed or appraised; (2) Base performance evaluations on specific predetermined job requirements or standards; (3) Do not rate people in any particular order (i.e., don't rate the best or the worst people first); (4) Rate people on the extent to which they fulfill the requirements of the job; compare people after, not before, an evaluation; (5) Avoid appraisal scales with vague benchmarks, such as "excellent," "above average," and so on. Use scales on which one merely records the frequency with which a behavior has been observed or on which the benchmarks themselves are defined behaviorally.

Fourth, to minimize the effect of first impression, there are two possible solutions: (1) Reserve all judgments about an teacher until the end of the time period for which the appraisal is scheduled; (2) Be a note taker rather than an evaluator during the interval between performance appraisals. Ideally, supervisors should record daily a subordinate's observed behaviors that lead to adequate or inadequate performance on job assignments. The incidents should be reviewed later by the manager when it is time assign ratings. Read the incidents in an order other than the recorded sequence.

Finally, to diminish the positive and negative leniency error, raters need to be trained to record exactly what they saw and to compare what they recorded with critical job behaviors/standards required in a job description or contained in the appraisal instrument.

Implementation of these training suggestions to minimize rating errors needs to be examined based on long-term effectiveness because appraisal abilities are not only characterized by knowledge of job performance as well as adequate judgments, but also by rating behavior and rating concepts. The effectiveness of changing a manager's concept or behavior associated with rating can not be seen immediately. Besides that, possible solutions to rating errors need to be assessed to demonstrate that specific

solutions are hitting the right points of individual rating error.

Conclusions

Personnel evaluation, which can be regarded as a process of judgment, has as its main purpose to increase teacher productivity. It may be conducted formatively or summatively and may be subjective-oriented or objective-oriented, depending on how accurately and consistently appropriate systems of personnel evaluation are used. To achieve this objective, it is necessary to make certain that an evaluation achieves the desirable reliability and validity.

Reliability and validity should be taken into account in the matter of the whole consideration of evaluation accuracy. Reliability represents the degree of stability and consistency of an evaluation, and should be a dominant premise of validity, which represents the degree of accuracy of an evaluation.

Evaluation errors come from many sources, such as the purposes of appraisal, teachers' performance, instrumentation dimensional differences, inter-rater differences, and, for most rating errors, from intra-rater bias. Some training programs for appraisers have been examined and found to effectively minimize intra-rater errors such as the contrast error, the similar-to-me error, the first impression error, the leniency error, the halo effect error, and so on. Even though they lack long-term investigation on the effectiveness of evaluation, appraisers can be expected to improve their appraising abilities, identify sources of errors, and avoid many possible rating errors through lectures to increase knowledge of appraisal theories as well as possible errors during the rating processes, through group discussions, and through being given feedback in some simulative appraisal practice.

References

Austin, J. T., & Villanova, P. (1992). The criterion problem: 1917-1992. *Journal of Applied Psychology, 77,* 836-874.

Borman, W. C. (1979). Individual differences correlations of accuracy in evaluating others performance effectiveness. *Applied Psychological Measurement, 3,* 103-115. Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Cronbach, L. J., & Gleser, G. C. (1957). *Psychological tests and personnel decisions.* Urbana, IL: University of Illinois.

Dickson, T. D., Hedge, J. W., & Teachout, M. S. (1987, August). *Accuracy of performance ratings: A structural model of rater attributes.* Paper presented at the annual meeting of the American Psychological Association, New York, NY.

Dwyer, C. A., & Villegas, A. M. (1993). *Guiding conceptions and principles for the Praxis Series: Professional assessments for beginning teachers* (Eric Document Reproduction Service no. ED385594)

Hedge, J. W., & Laue, F. J. (1988, August). *Can appraisers rate work performance accurately?* Paper presented at the annual meeting of the American Psychological Association, Atlanta, GA.

Hopkins, K. D. (1998). *Educational and psychological measurement and evaluation.* Needham Heights, MA: Allyn & Bacon.

Landy, F. J., & Farr, J. L. (1975). Performance rating. *Psychological Bulletin, 87,* 72-107.

Latham, G. P., & Wexley, K. N. (1994). *Increasing productivity through performance appraisal.* Menlo Park, CA: Addison-Wesley.

Latham, G. P., Wexley, K. N., & Pursell, E. D. (1975). Training managers to minimize rating errors in the observation of behavior. *Journal of Applied Psychology, 60,* 550-555.

Linn, R. B., & Gronlund, N. E. (2000). Measurement and assessment in teaching (8thed.). Upper Saddle River, NJ: Prentice Hall.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). Upper Saddle River, NJ: Prentice Hall.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational measurement: Issue and practice, 14*(4), 5-8.

Peterson, K. D. (1995). *Teacher evaluation: A comprehensive guide to new directions and practices.* Thousand Oaks, CA: Corwin.

Taft, R. (1955). The ability to judge people. *Psychological Bulletin, 52,* 1-23.

Weitz, J. (1961). Criteria for criteria. *American Psychologist, 16,* 228-231.