# RELIABILITY AND VALIDITY OF PSYCHOMETRIC INSTRUMENTS: THE EXAMPLE OF STUDENT EVALUATION OF TEACHING EFFECTIVENESS

## INTRODUCTION

Psychometric questionnaires are based upon the assumption that the answers given and the consequent scores gained are accurate measures of behaviour. In order to achieve this, such instruments undergo rigorous reliability and validity establishing.

This article outlines the main types of reliability and validity for psychometric instruments by looking at the example of student evaluation of teaching effectiveness (SETE). In the USA, in particular, students are asked to rate their lecturers in terms of quality of teaching, for example. Psychometric questionnaires used to do this have to show that the scores gained for lecturers are an accurate measure of teaching effectiveness rather than just the students' unsubstantiated opinions. This is the establishing of validity (ie the questionnaire measures what it claims to measure). While reliability is the consistency of the questionnaire.

Theall and Franklin (2001) are confident about students' ratings: "No one else is as qualified to report on what transpired during the term simply because no one else is present for as much of the term" (p48).

Marsh (1984) outlined the purpose of student evaluation of teaching as fourfold:

i) As a "diagnostic feedback to faculty about the effectiveness of their teaching" (p707);

ii) For use in teacher promotion decisions;

iii) As information for students in selection of courses;

iv) To show the "outcome on a process description for research or teaching" (p707).

## RELIABILITY OF STUDENT RATINGS

Here reliability refers to the fact that the ratings will measure the same score every time, ie the same lecturer producing the same quality lecture on two occasions will receive the same rating by the same student.

Table 9 shows the types of reliability for psychometric questionnaires.

| TYPE OF RELIABILITY | DESCRIPTION |
|---|---|
| Internal<br><br>1. Split-halves<br>a. odd and even questions<br>b. all possible split-halves<br><br>2. Parallel/multiple forms<br>a. with time interval<br>b. without time interval<br><br>3. Item analysis | <br><br>- Correlation between scores on two halves of test<br><br>- Correlation between scores on two versions of same test<br><br>- Ability of each item to discriminate between high and low scorers |
| External<br><br>Test-retest<br>- immediate or later | <br><br>- Correlation of scores between same test repeated at different times |

Table 9 - Different types of reliability for psychometric instruments.

Doyle (1975) listed the sources of reliability errors:

i) Computational error - eg putting the wrong instructor's name on ratings summary.

ii) Rater's task - ie problem with nature of the questions used.

iii) Environment - physical or social environment.

iv) Rater - lacks motivation or memory problems, as well as:

- Halo effect: overall impression influences specific rating items.
- Leniency error: tendency to rate higher when known that ratings being used for promotion purposes.
- Central tendency: inclination for mid-point on scale.
- Proximity error: rate adjacent items similarly.
- Contrast error: projection of own deficiencies on to ratee.
- Logical error: rating traits that "ought" to go together.

The first study of reliability came from Guthrie (1927). Two hundred and eighty-five psychology students ranked lecturers at the University of Washington, and

then again two weeks later. A correlation of r = 0.89 was found.

In Britain, Foy (1969) followed up his study with Cooper (Cooper and Foy 1967), due to objections about the original findings on an ideal lecturer. A different group of students used the same check-list as the first study, and there was a correlation of 0.93 between the two ratings (1 in 2000 possibility of a chance correlation as high as that). This seems the most straightforward evidence of the reliability of an instrument.

## Methods Used to Establish Reliability

### 1. Internal Consistency

The aim is to correlate various questions within the instrument. Using for example, odd-even or split-half, and coefficient alpha (Cronbach 1951) or Kuder-Richardson formulas (Kuder and Richardson 1937).

Feldman (1977) reports an extension of this approach, where two mean scores for a particular item can be obtained by randomly dividing a class in half. Most of the commonly used instruments report reliability coefficients over 0.50.

### 2. Test-Retest

Here the rating instrument is given to the same subjects at two different times. The aim being to correlate the two scores of each subject.

But the instructor may change between administrations of the instrument, and so a small correlation will suggest that the instrument is unstable. This method is also criticised for "being a test of the student's memory instead of being a measure of reliability" (Frey 1978 p85).

### 3. Mean Ratings

It is assumed that mean ratings of instructors should be different, because the instructors display different teaching behaviour. If the means are similar or identical, the ratings are seen as biased.

But the assumption that instructors do differ is open to question.

Frey (1978) used a variation of this method. He chose a sample of the data representing instructors who had taught three or more classes (with 10 + students in

each), which had filled in ratings. Variance estimates were calculated for differences among instructors, and differences among classes within instructors - inter rater agreement. A formula used showed the proportion of observed variance due to differences in instructor.

4. Analysis of Variance (ANOVA)

Proposed by Guilford (1954): rather than attempting to remove potential bias, it aims to identify the contribution of bias to the final rating, and adjust for it. Obviously, this has advantages because some potential biases cannot be easily separated, like the halo effect.

Treffinger and Feldhusen (1970), using this method, found that the halo effect only accounted for 10% of the variance in students' ratings (quoted in Doyle 1975 p43).

5.Inter-Rater Reliability

This looks at the consistency of ratings among people. Reliability here is when all raters in a group give the same pattern of responses. Usually estimated by intraclass correlation coefficients, ie the comparison of ratings within one class of one lecturer with ratings of different instructors. Because it is sensitive to the number of raters, Centra (1979) suggested intraclass correlations of .70s for ten raters through to .90s for twenty.

Feldman (1977) makes a number of points about interpreting the reliability coefficients:

i) "reliability coefficients of individual ratings indicate the degree of general or relative consistency among raters; they do not measure exact or absolute agreement" (p229);

ii) inter-rater agreement is only the degree to which independent raters give the same rating for the same lecturer;

iii) inter-rater reliablity is "the degree to which the ratings by different raters are proportional when expressed as deviations from their means" (p229);

iv) the reliability coefficients of average college student ratings may be high, but this does not mean that individual students within the classes are highly consistent in their ratings;

v) consistency in ratings among students may not be a good basis for estimating individual ratings or average

ratings reliability, particularly if the aim is to compare ratings across situations.

Guthrie (1927) suggested that student ratings agree at the end of the term because of greater exposure to the lecturer, or student gossip.

Overall, establishing reliability of a psychometric questionnaire is probably easier than establishing validity.

## VALIDITY OF STUDENT RATINGS

Do students know a good lecturer, ie are student ratings actually measuring good teaching? Here validity means that the ratings are an accurate assessment of teaching quality, not other factors, like class size or personality of student. There are different types of validity (table 10).

| TYPES OF VALIDITY | DESCRIPTION |
|---|---|
| Face | Based on commonsense; the items appear valid |
| Content | Sophisticated version of face validity; experts see the items as valid |
| Criterion<br><br>a. Predictive<br><br>b. Concurrent | <br><br>a. Correlation of test score with future performance<br>b. Correlation of test score with another test of same thing |
| Construct | Correlation of test score with behavioural measure of same thing |
| Discriminant | Correlation of test score with different measures of same behaviour (some expected and some unexpected); extension of construct validity |

Table 10 - Different types of validity for psychometric questionnaires.

McBean and Al-Nassri (1982) noted that "students strongly believed that student evaluations do measure teacher effectiveness ... while faculty only slightly agreed" (p278). This statement can be said to show face

validity. Some would argue, though, that this is only valid as an indicator of student satisfaction.

## Criterion Validity

This concentrates on the relationship of ratings with other objective measures. The most common measure used is student learning (usually defined as the grade in the course examination).

In a now famous study in "Science", Rodin and Rodin (1972) found a negative correlation between the amount learned from classes, and their rating of the teacher. The Rodins used a subjective rating of the lecturer, and an objective measure of the amount of calculus learned. The conclusion of r = -0.75 correlation threatened the validity of students' evaluation ratings.

But subsequent studies have consistently found positive correlations. Frey (1978) outlined a number of problems with the Rodins study - for example, study based on teaching assistants rather than teachers who gave the main lectures. Further on in his article, after reviewing the studies since Rodins, Frey pointed out the need to study the "regular instructors", and to use " a rating form which emphasises the appropriate teaching traits" (p75).

Frey (1978) in testing the validity of the two dimensions of "skill" and "rapport" of the Instructional Rating Card (Frey et al 1975), correlated each with examination scores. Using a course divided into multiple sections, taught by different instructors, but with a common syllabus, textbook, and examination. The median correlations were different: for the "skill" factor, it was r = 0.81 but for "rapport" it was r = 0.29. "The two rating factors are clearly not the same in their ability to indicate which teachers were most effective in preparing their students for the final examination" (p87) [6].

Doyle (1983) has his problems with using a student achievement test as the criterion for establishing the validity of student ratings of instruction:

i) some characteristics of teaching are not linked to test scores - eg "clarity" and "rapport";

ii) it is assumed that the relationship is a linear one and thus the Pearson product-moment correlation can

[6] More recently, Spooren and Mortelmans (2006) have identified an underlying factor called "teacher professionalism".

be used. But it is possible that it is a non-linear relationship between student achievement and student ratings of instruction;

iii) which unit of analysis should be used:

a) pooled within-class analysis (individual ratings in each section of the course, and average across course);
b) between-sections analysis (mean ratings of evaluation items across course);
c) total-class approach (individual ratings).
Doyle prefers the first approach;

iv) if participants are randomly divided into sections of the course, then the generalizability of findings are limited.

Emery et al (2003) revisited many of these problems and others.

The main alternative to final grade is to use students' gains in knowledge. But there are problems in how to measure the gain.
Marsh and Overall (1980) tried to combine both criteria. They used final examination grade, ability to apply course material, and inclination to pursue the subject further. The first is seen as a cognitive criterion, while the other two are self-reported affective criteria. The students used were taking a course in computer programming. The authors, accepting methodological weaknesses, felt that more than one construct must be used to establish validity. "Therefore, because there is no universally accepted criterion of effective teaching, the validation of any teaching effectiveness measure must focus on a wide range of indicators" (p474).

Obviously, the higher the correlation, the better for validation. But validity will be specific to a particular situation, and "must always be evaluated in relation to a situation as similar as possible to the one in which the measure is to be used" (Thorndike and Hagen 1977 p69).

## Construct Validity

For some researchers, criterion validity is not a satisfactory method to establish the validity of student ratings of instruction because effective teaching is a construct. Thus for them construct validation is the best method.
The main aim is to correlate multiple indicators of

effective teaching. For example, student ratings and various criteria assessed for convergent and discriminant validity.

Howard et al (1985) used this method to establish teaching effectiveness using student ratings, colleagues ratings, teacher self-ratings, former-student ratings, and trained observers. Ratings by current and former students were most effective.

The main criteria used in construct validation are self-evaluation by the lecturer, colleagues' evaluation, external observers, administrators, former students' evaluations, and the research productivity of lecturers.

## 1.Lecturer Self-Rating

There is a general tendency for instructors to rate themselves more favourably than their students do. But there is agreement on instructor's strengths and weaknesses.

Centra (1972) found differences also between faculties: instructors in natural sciences rated effort needed for their course less than did the students, while education, business, home economics, and nursing instructors were the opposite.

Marsh (1982), quoting his own studies, found correlations of r = 0.41 for undergraduate ratings, and r = 0.39 for postgraduate ratings, with lecturer's self-evaluation.

## 2.Ratings by Colleagues

In their early literature review, Costin et al (1971) found correlations between 0.30 and 0.63 for students' ratings and colleagues' ratings.

But in most cases, colleagues' ratings are not based on sitting through the lecture, but on "student hearsay, on the observation of the presumed effects of instruction ... and on inferences from their personal acquaintances (with the colleagues)" (Guthrie 1949 p113).

Ballard, Reardon and Nelson (1976) found correlations that ranged from 0.62 to 0.84. Studies based on colleagues actual visitation to the classroom are limited.

Furthermore, there is the problem that the presence of an observer can change the classroom situation - for example, by effecting the performance of the lecturer.

Murray (1980) argued that peer ratings are "less sensitive, reliable and valid" (p45) than student ratings.

## 3.Observation by External Observers

Murray (1980) felt that student ratings "can be accurately predicted from outside observer reports of specific classroom teaching behaviours" (p31). The feeling is that trained observers are best, and particularly if they concentrate on specific behaviour (eg clarity-related behaviour: number of false starts or halts in speech, redundantly spoken words, and tangles in words) (Marsh 1984).

## 4.Administrators' View

Cotsonas and Kaiser (1962) used clinical students in a medical school, and compared their ratings with departmental administrators. The former tended to stress the attitude towards students, and teaching skill, while the latter stressed knowledge. The authors suggested that the administrators noted the knowledge of the lecturer, and then assumed the other abilities ("halo effect"). It would also seem that the administrators took into account more than just classroom behaviour, but also their general judgments about the lecturer.

## 5.Retrospective Ratings of Alumni

Graduating students were asked to nominate "most outstanding" and "least outstanding" lecturers in their departments. Then undergraduates were asked to rate the nominated lecturers. Results indicated that the "most outstanding" lecturers were rated higher than the "least outstanding". A correlation of r = 0.82 between graduates' and undergraduates' choices of most and least outstanding (Marsh 1977).

Gaski (1987) urged caution when using former students' ratings for validity purposes because "the similarity between the student and former student teaching evaluations can be explained if the primary determinant of the former student ratings is former students' recollection of the assessment they made when they were current students of the given instructor one or two years earlier" (p329).

## 6.Research Productivity

Blackburn (1974) suggested research and effective teaching were opposites. For example, McDaniel and Feldhusen (1970) found significant a negative correlation between first authorship of books and students' ratings of teaching. But a significant positive correlation

between second authorship of professional articles and rating of teaching.

Marsh (1984) finds no correlation or a small positive correlation between the two. "Although these findings seem to neither support nor refute the validity of student ratings, they do demonstrate that measures of research productivity cannot be used to infer teaching effectiveness or vice versa" (p729).

Feldman (1989) undertook a detailed literature review of the North American studies comparing overall ratings of teaching effectiveness made by current and former students, lecturers' colleagues, administrators, external (neutral) observers, and teachers' self-evaluation. The results are summarised in table 11.

Feldman concluded that there was similarity between various raters, in this order: current students and colleagues; current students and administrators; colleagues and administrators (similar in relative assessment, but not in absolute assessment); self-evaluation and current students; self-evaluation and colleagues. For the other relationships, there were not enough studies to determine.

Berk (2005), more recently, extended this type of analysis using twelve sources of evidence.

| Method Used | Current Students | Former Students | External Observers | Colleague | Adminis-trators |
|---|---|---|---|---|---|
| Current Students | | +.69(6)* | +.50(5)* | +.55(14)* | +.39(11)* |
| Former Students | | | +.08(1) | +.33(1) | no cases |
| External Observers | | | | -.12(1) | no cases |
| Colleague | | | | | +.48(5)* |
| Administrators | | | | | |

(* = significant correlation p<0.001 two-tailed. The number in () is number of studies found)

Table 11 - Summary of the studies found by Feldman (1989) showing a correlation between different methods of assessing teaching effectiveness.

## Use of MTMM

A number of criteria are used under the heading of the Multi-Trait Multi-Method (MTMM) approach (Campbell and Fiske 1959). The use of a number of methods to measure one trait/construct allows correlations to be

made; thus producing a MTMM matrix. It allows the estimation of variance due to traits or methods, and of unique or error variance.

It is possible to show convergent validity (correlation between items that should go together) and divergent validity (small or no correlation between items that should not go together). This method allows the research to estimate the effects of bias; for example, method bias (large correlation between variables because of the method used).

Murphy and Davidshofer (1988) summarised three points that a test will possess as established effectively by MTMM:

> 1.Scores on the test will be consistent with scores obtained using other measures of the same construct.
> 2.The test will yield scores that are not correlated with measures that are theoretically unrelated to the construct being measured.
> 3. The method of measurement employed by the test shows little evidence of bias (p106).

In their original article, Campbell and Fiske proposed a series of rules to follow for evaluating convergent and discriminant validity:

> 1. The convergent validity coefficients should be statistically significant and sufficiently different from zero to warrant further examination of the validity.
> 2. The convergent validities should be higher than correlations between different traits assessed by different methods.
> 3. The convergent validities should be higher than correlations between different traits assessed by the same method.
> 4. The pattern of correlations between different traits should be similar for each of the different methods (quoted in Marsh and Hocevar 1983 p233).

The above rules have been criticised. Firstly, over what constitutes a satisfactory result.

Secondly, the use of correlations based on observed variables to draw conclusions about underlying factors (Kenny and Kashy 1992).

Attempts have been made to establish validity by using large multi-section courses, where different groups of students are presented the same material by different instructors.

Ideally the following controls should be used:

- many sections to the course;

- random assignment of students to the sections;
- pre-test measures used;
- each section taught by separate instructors;
- the final examination graded externally;
- common textbooks among the sections (Marsh 1984).

Validity is then assessed by correlating the student ratings in each section.

## CONCLUSIONS

Centra (2003) believed that SETE instruments are reliable and stable, and valid when compared with student learning.

The question of establishing validity has become a methodological issue debated in the literature, particularly around the use of criterion validity (established through multi-section courses) or construct validity (established using MTMM).

However, taking into account the weaknesses of the use of the different criteria, it is fair to say that student ratings of instruction are valid. But the criteria used are validity measures of what?

Feldman (1977) looked at the purpose of the ratings - if it is to obtain objective descriptions of teachers, there may be a problem, but not if it is to measure students' subjective responses.

## REFERENCES

Ballard, M; Reardon, J & Nelson, J (1976) Student and peer rating of faculty Teaching of Psychology 3, 88-90

Berk, R.A (2005) Survey of twelve strategies to measure teaching effectiveness International Journal of Teaching and Learning in Higher Education 17, 1, 48-62

Blackburn, R.T (1974) The meaning of work in academia. In Doi, J (ed) Assessing Faculty Effort San Francisco: Jossey Bass

Campbell, D.T & Fiske, D.W (1959) Convergent and discriminant validation by the MTMM matrix Psychological Bulletin 56, 81-105

Centra, J.A (1972) Two Studies on the Utility of Student Ratings for Improving Teaching SIR Report no.2; Princeton, NJ: Educational Testing Service

Centra, J.A (1974) The relationship between student

and alumni ratings of teachers Educational and Psychological Measurement 34, 321-325

Centra, J.A (2003) Will teachers receive higher student evaluations by giving higher grades and less course work? Research in Higher Education 44, 5, 495-518

Cohen, P.A (1981) Student ratings of instruction and student achievement: a meta-analysis of multisection validity studies Review of Educational Studies 51, 3, 281-309

Cooper, B & Foy, J (1967) Evaluating the effectiveness of lectures Universities Quarterly 21, 2, 182-185

Costin, F; Greenough, W.T & Menges, R.J (1971) Student ratings of college teaching: reliability, validity, and usefulness Review of Educational Research 41, 511-535

Cotsonas, N.J & Kaiser, H.F (1962) A factor analysis of students' and administrators' ratings of clinical teachers in a medical school Journal of Educational Psychology 53, 219-223

Cronbach, L.J (1951) Coefficient alpha and the internal structure of tests Psychometrika 16, 297-334

Doyle, K.O (1975) Student Evaluation of Instruction Lexington, Mass: Lexington Books

Doyle, K.O (1983) Evaluating Teaching Lexington, Mass: Lexington Books

Emery, C.R; Kramer, T.R & Tian, R.G (2003) Return to academic standards: A critique of student evaluation of teaching effectiveness Quality Assurance in Education 11, 1, 37-46

Feldman, K.A (1977) Consistency and variability among college students in rating their teachers and courses: a review and analysis Research in Higher Education 6, 223-274

Feldman, K.A (1989) Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators, and external (neutral) observers Research in Higher Education 30, 2, 137-194

Foy, J (1969) A note on lecturer evaluation by students Universities Quarterly 23, 3, 345-349

Frey, P.W (1978) A two-dimensional analysis of

student ratings of instruction Research in Higher Education 9, 69-91

Frey, P.W; Leonard, D.W & Beatty, W.W (1975) Student ratings on instruction: Validation research American Educational Research Journal 12, 435-447

Gaski, J.F (1987) On "construct validity of measures of college teaching effectiveness" Journal of Educational Psychology 79, 3, 326-330

Guilford, J.P (1954) Psychometric Methods (2nd ed) New York: McGraw-Hill

Guthrie, E.R (1927) Measuring student opinion of teachers School and Society 25, 175-176

Guthrie, E.R (1949) The evaluation of teaching Educational Record 30, 109-115

Howard, G.S; Conway, G.C & Maxwell, S.E (1985) Construct validity measures of college teaching effectiveness Journal of Educational Psychology 77, 2, 187-196

Kenny, D.A & Kashy, D.A (1992) Analysis of the MTMM matrix by confirmatory factor analysis Psychological Bulletin 112, 1, 165-172

Kuder, G.F & Richardson, M.W (1937) The theory of the estimation of test reliability Psychometrika 2, 151-160

McBean, E.A & Al-Nassri, S (1982) Questionnaire design for student measurement of teaching effectiveness Higher Education 11, 273-288

McDaniel, E.D & Feldhusen, J.F (1970) Relationships between faculty ratings and indexes of service and scholarship Proceedings of the 78th Annual Convention of the American Psychological Association 5, 619-620

Marsh, H.W (1977) The validity of students' evaluations: classroom evaluations of instructors independently nominated as best and worst teachers by graduating seniors American Educational Research Journal 14, 4, 441-447

Marsh, H.W (1982) SEEQ: a reliable, valid and useful instrument for collecting students' evaluations of university teaching British Journal of Educational Psychology B52, 77-95

Marsh, H.W (1984) Students' evaluations of university teaching: dimensionality, reliability,

validity, potential biases, and utility Journal of Educational Psychology 76, 5, 707-754

Marsh, H.W (1987) Students' evaluations of university teaching: research findings, methodological issues, and directions to future research International Journal of Educational Research 11, 253-388

Marsh, H.W & Hocevar, D (1983) Confirmatory factor analysis of MTMM matrices Journal of Educational Measurement 20, 3, 231-248

Marsh, H.W & Overall, J.U (1980) Validity of students' evaluations of teaching effectiveness: cognitive and affective criteria Journal of Educational Psychology 72, 4, 468-475

Murphy, K.R & Davidshofer, C.O (1988) Psychological Testing: Principles and Applications Englewood Cliffs, NJ: Prentice Hall

Murray, H.G (1980) Evaluating University Teaching: A Review of Research Toronto: Ontario Confederation of University Faculty Associations

Rodin, M & Rodin, B (1972) Student evaluation of teachers Science 177, 1164-1166

Spooren, P & Mortelmans, D (2005) Teacher professionalism and student evaluation of teaching: Will better teachers receive higher ratings and will better students give higher ratings? Educational Studies 32, 2, 201-214

Theall, M & Franklin, J (2001) Looking for bias in all the wrong places: A search for truth or a witch hunt in student ratings of instruction? New Directions in Institutional Research 27, 5, 45-56

Thorndike, R.L & Hagen, E (1977) Measurement and Evaluation in Psychology and Education (4th ed) New York: John Wiley