Validity of Student Evaluations: A Comparison of Student Ratings to Instructor Self-Evaluation

1987

Rebecca A. Hopson
University of Central Florida

Find similar works at: https://stars.library.ucf.edu/rtd

University of Central Florida Libraries http://library.ucf.edu

Part of the Industrial and Organizational Psychology Commons

STARS Citation

https://stars.library.ucf.edu/rtd/5016

This Doctoral Dissertation (Open Access) is brought to you for free and open access by STARS. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of STARS. For more information, please contact lee.dotson@ucf.edu.
VALIDITY OF STUDENT EVALUATIONS: A COMPARISON OF STUDENT RATINGS TO INSTRUCTOR SELF-EVALUATIONS

BY

REBECCA ANNE HOPSON
B.S., University of Central Florida, 1979

THESIS

Submitted in partial fulfillment of the requirements for the Master of Science degree in Industrial/Organizational Psychology in the Graduate Studies Program of the College of Arts and Sciences University of Central Florida Orlando, Florida

Summer Term
1987
TABLE OF CONTENTS

LIST OF TABLES ........................................ iii

LITERATURE REVIEW ..................................... 1

METHOD .................................................. 10
  Subjects .............................................. 10
  Instrument .......................................... 10
  Procedure .......................................... 10

RESULTS .................................................. 12

DISCUSSION ............................................. 18

APPENDICES ............................................ 21
  A. Request to Human Subjects Committee ............ 22
  B. "Nature of Study" Letter to Participants ......... 23
  C. Informed Consent Agreement ........................ 24
  D. Instructor Evaluation Form ........................ 25
  E. Instructor Self-Evaluation Form ................. 28

REFERENCES ............................................ 31
## LIST OF TABLES

1. Seven Evaluation Form Factors and Items Within Each .................................. 14
2. Means and Standard Deviations of Instructors Self- and Student Evaluations .................. 15
3. Convergent and Discriminant Validity of Student and Instructor Evaluations for 14 Courses ........ 16
LITERATURE REVIEW

Student evaluations of teaching instruction are commonly used in many colleges and universities for a variety of purposes. The two most prominent usages are to provide diagnostic feedback to college instructors regarding the effectiveness of their teaching and to provide a measure of teaching effectiveness to be considered in tenure/promotional decisions. The extent to which student ratings are used in personnel decision-making varies for many institutions. Marsh (1984) reports that some universities require systematic student input before making promotional decisions, while others consider it to be optional. He also concludes, based on an examination of a variety of surveys, that the use of student evaluation of teaching effectiveness has increased over the last 25 years.

While student evaluations are one of the most ubiquitous evaluation sources of college instruction, they are also the most controversial. Many instructors continue to question their relative usefulness, particularly when such ratings are used in personnel decisions. Marsh (1982b) reported that while 80% of faculty agreed that student evaluations are useful to them as feedback, only 38% felt that these ratings gave an accurate assessment of instructional quality.

One major concern is that student ratings do not accurately reflect effective teaching. Particular characteristics of the instructor and/or student are purported to affect the ratings
given. Numerous studies have investigated such factors as sex of instructor, class size, course context, workload, and grade point average as possible sources of contamination. Review of the literature on student evaluations and the effects of extraneous variables have yielded inconsistent results (Marsh, 1984). Similar conclusions have been drawn by McKeachie (1979) and Centra (1977). The conflicting results have increased university instructors consensus regarding the validity and reliability of student ratings used in the assessment of their teaching proficiency.

Much of the controversy and difficulty involved in validating student ratings also stems from the lack of a single criterion to assess effective teaching. Previous research has validated student ratings against a variety of different criteria. The most commonly used criterion is student performance on a standardized test. Studies have generally been limited to large multisection courses in which different instructors present the same materials to different groups of students. Several researchers found that when different instructors taught different sections of the same course, the sections that performed better on the examination rated their instructors higher than did lower performers (Centra, 1977; Frey, Leonard, and Beatty, 1975; Marsh, Fliner, and Thomas, 1975; and Marsh and Overall, 1980).

While the use of student learning as a criterion has supported the validity of student ratings, research has generally been limited to specialized settings (i.e., multisection courses). Critics have argued that such a criterion would be very difficult to assess across a wide range of courses which normally cannot
be compared, thereby limiting the generality of this type of setting. Firth (1979) found similar agreement between the two sets of ratings when collecting student evaluations at the end of graduation from college and one year later.

Despite the support various criteria have given to the validity of student ratings, skeptics have yet to be convinced. Marsh (1982) contends that unless a suitable criterion is utilized which is both applicable across a wide variety of courses and is acceptable to faculty, the validity of student ratings will continue to be questioned. He advocated the use of instructor self-evaluations as the one criterion that meets both of these requirements.

Much of the research that has been conducted using instructor self-evaluations as the criterion by which to validate student ratings has sought to convince faculty and administration of its relative worth. Previous researchers have attempted to eliminate many of the criticisms of student ratings by avoiding the use of particular criteria that are either course-specific (e.g., standardized tests in multisection courses) or that can be regarded as inappropriate (e.g., ratings by former students). Using instructor self-evaluations as the criterion by which to validate student ratings may be a more acceptable method of assessment.

Studies that have used instructor self-evaluations as the criterion by which to validate student evaluations have yielded mixed results. Blackburn and Clarke (1975) examined the correlates between administrator, colleague, student, and self-ratings at a liberal arts college. Forty-five faculty members were rated on both
teaching effectiveness and overall contribution to the college by each of the four rating methods. They reported near-zero correlates between instructor self-ratings and each of the other three criteria. The authors concluded that for student, administrator, and colleague-ratings, considerable variation in the factors entered into performance judgements. The data indicated only a slight agreement between self-ratings and faculty colleagues on the same performance dimension and almost no relationship with judgements made by administrators. Student and self-ratings yielded a correlation of .19. As Marsh (1984) points out, this low correlation between student and self-ratings could have in part, been influenced by the fact that faculty self-evaluations were only general impressions of teaching effectiveness while student evaluations were based on actual teaching behaviors in a single course.

Centra (1973) conducted a study, whereby faculty members were asked to select one particular course in which to base their self-ratings and be rated by students. The data revealed a coefficient of .21 which was consistent to that found in the Blackburn/Clark study. Again, Marsh (1984) points out that the methodologies used and low reliability of the measuring instruments could in part explain the low coefficient findings.

In another small study with ratings of fewer than 20 instructors, Braskamp, Caulley, and Costing (1979) investigated the interrelationships among instructor self-ratings, student ratings, and student achievement at a large midwestern university. The researchers compared self and student ratings to student achievement using 17 introductory psychology courses over a two-semester period. The researchers used a multitrait-multimethod matrix to demonstrate convergent and
discriminant validity. The findings revealed good convergent validity between instructor self-evaluations and student ratings during the second semester. They reported correlations of .31 for the first semester and .61 for the second. They also reported significant discriminant validity with higher intercorrelations for student ratings than instructors. Similarly, the data revealed very little discriminant validity the first semester yet the second semester demonstrated sufficient discriminant validity on those factors which yielded the highest convergent validities.

Braskamp et al. (1979) suggested that the low correlations found in the first semester may have been a result of the fact that the instructors used in these studies were relatively inexperienced. During the initial semester they may have been basing much of their self-image as a teacher on student opinions as their average self-ratings as a group increased the second semester. These findings lead the researchers to conclude that early in their career, instructors must have little other information to judge their performance than from that of the students' opinions.

Doyle and Crichton (1978) examined the convergent and discriminant validity of student ratings, peer ratings, and self-ratings of college instruction utilizing a multitrait-multimethod matrix. The researchers found that self, student, and colleagues ratings were somewhat similar in mean, range, distribution, and skew, with students giving the least favorable ratings and peers most favorable. The student, colleague and self-ratings revealed better discriminant validity than colleague ratings. Doyle and Crichton (1978) reported a median correlation of .47 for student and self-ratings.
Using a somewhat larger sample than the previously mentioned studies, Webb and Nolan (1955) examined the relation between student, supervisor, and self-ratings. Fifty-one instructors in a military setting rated themselves on the same teaching proficiency scale as their students and colleagues. The data indicated that student ratings and the self-ratings of the instructors were in high agreement with a correlation of .62.

Marsh, Overall, and Kesler (1979) made a comparison of faculty self-evaluations and student ratings to provide further insight into the validity of student evaluations of teaching effectiveness. Using separate factor analyses, they demonstrated that the same evaluation factors underlay both student and faculty evaluations, and the median correlation between student and instructor ratings was .49. Marsh et al. (1979) concluded that their findings reaffirmed the validity of student evaluations and also indicated that self-ratings can be a useful source of evidence.

The Marsh (1982b) study was based on the preceding research findings of Marsh, Overall, and Kesler (1979). Although a replication of the earlier one, this study differed in several ways. First, new dimensions were added to the survey instrument. Second, the study included teaching assistants and graduate level courses which were not included in the earlier study. Third, the sample size was increased to include 329 courses. As was demonstrated in the Marsh et al. (1979) study, separate factor analyses of teacher and student responses identified the same evaluation factors with a median correlation of .45. Using a multitrait-multimethod analysis, Marsh (1982b) provided support for both convergent and divergent
validity of the rating factors. In agreement with the earlier study, Marsh (1982b) concluded that the findings supported the validity of student ratings and further emphasized the importance of employing multifactor evaluation instruments that are developed with the use of factor analysis.

A considerable body of empirical research has provided clear support for the multidimensionality of student evaluations. The assumption that student ratings, like the teaching they represent, reflect multiple dimensions, is based on a logical analysis of the content of effective teaching. Evaluation instruments that demonstrate content validity have well-defined factor structure and provide measures of distinct components of teaching proficiency.

Two examples of such evaluation instruments are Marsh's Students' Evaluations of Educational Quality (SEEQ) (Marsh, 1982a, 1982b, 1983) and Frey's Endeavor instrument (Frey et al., 1975). Separate factor analyses of responses to each of these instruments demonstrated that student evaluations of teaching effectiveness do measure distinct components. The Marsh (1982b) study asked 329 instructors to evaluate their own teaching proficiency using the same SEEQ instrument as students. Not only did this study demonstrate the multidimensionality of student ratings but also found that similar ratings underlie faculty evaluations of their own teaching effectiveness.

If an evaluation instrument is composed of distinguishable components of effective teaching, and more than one evaluative source is utilized, multitrait-multimethod analyses can be used to permit meaningful interpretation of what is being measured. The design of instruments should include separate components to reflect the
Inherent multidimensionality of teaching effectiveness. A survey instrument that includes different measures can demonstrate that the items within the same group do measure distinct components of teaching effectiveness (Marsh, 1984). Most of the research that has compared student ratings to instructor self-ratings has incorporated more than one trait and more than one method of evaluation into the assessment of validity. When instructors are asked to evaluate themselves using the same survey instrument as their students, the extent of agreement on each measure can be determined, as well as the uniqueness of each dimension. Convergent validity can be demonstrated if student and instructor ratings on the same evaluation dimensions are correlated (Marsh, 1982). Convergent validity determines the extent to which different groups of raters agree on their ratings of each scale. If such agreement is demonstrated by a significant correlation, convergent validity has been demonstrated.

Discriminant validity or divergent validity demonstrates the distinctiveness of the various rating items. Testing the specific validity of the different rating factors, as well as the ratings in general, can provide support for the uniqueness of each. Evidence of divergent validity argues in favor of using multifactor instruments and discourages the use of averages across a number of evaluative items and/or using a single overall rating.

In the present study, two major issues regarding the validity of student and instructor self-ratings will be considered. First, correlations between the same evaluation factors rated by the two groups will be examined to determine the degree of convergent validity. Secondly, discriminant validity will be determined by examining
whether student-faculty agreement on each factor is independent of agreement on other factors.

A set of 22 evaluation items were designed to measure the following 7 factors: Communication, Enthusiasm, Organization/Planning, Group Interaction, Individual Rapport, Subject Coverage, and Exams/Assignments. The multiple traits used in this study are the seven evaluation factors, and the multiple methods are students rating their instructors and instructors rating themselves.

It is hypothesized that student and instructor self-ratings will correlate significantly, thereby providing clear support for the validity of each. Secondly, it is hypothesized that discriminant validity will be demonstrated for each of the 7 evaluative factors.
METHOD

Subjects

Student evaluations and instructor self-evaluations were collected during the Fall 1986 semester at the University of Central Florida in undergraduate courses taught in the College of Arts and Sciences. Evaluation instruments were completed by 292 students and 14 college instructors.

Instrument

The evaluation instrument consisted of 22 evaluation items adapted from Students' Evaluation of Educational Quality/SEEQ, Frey's Endeavor Instrument, and U.C.F.'s Faculty Evaluation Form. While this form was derived from the above evaluation forms, the items were changed to a behavioral format and a 5-point frequency scale was utilized.

Procedure

Evaluation forms were administered at the end of the semester by the researcher. Students were informed that the data generated from this project would be used for a graduate thesis project and each were given a description of the nature of this study (see Appendix A). All subjects were informed that confidentiality would be maintained and that individual identification would not be requested.

University instructors were asked to evaluate their own teaching proficiency using the same rating instrument completed by students.
The only difference was that items were worded in first person. They were also given a description of the nature of this study and informed that the information obtained would be used for a graduate theses project.
RESULTS

In order to evaluate the hypothesis, a multitrait-multimethod analysis was used to first determine the degree of convergent validity or the extent to which different groups of raters agree on their ratings of each scale. Correlations between the same evaluation factors rated by the two different groups were examined to determine the degree of convergent validity. Secondly, discriminant validity was examined by determining whether student–faculty agreement on each factor was independent of agreement on other factors.

The coefficient alpha was used to obtain reliability coefficients for both groups of raters. Individual item scores were used to calculate the sum of item variances for each factor. A high reliability coefficient indicates that items within a factor have high intercorrelations with each other and are measures of the same trait. Low intercorrelations will yield a low reliability coefficient which indicates that items within a factor are measuring different traits.

The convergent validity coefficients were calculated using the Spearman Correlation method. The raw data for each subject were transformed into averages of the items within each of the seven factors. Student averages for each factor were compared to their corresponding instructor's averages when calculating the convergent correlation coefficients.
In Table 3, the correlations between the 22 evaluation items as evaluated by the same group are depicted in the triangular matrices. The upper left triangular matrix illustrates the intercorrelations among student evaluations. The triangular matrix on the lower right contains the intercorrelations among instructors self-evaluation factors. The diagonal of each represents the reliability coefficients. The square matrix on the lower left illustrates the correlations between student evaluation factors and instructor self-evaluation factors. The diagonal represents the convergent validity coefficients between the same evaluation factors assessed by the multiple methods.

The correlations between student evaluation factors and faculty self-evaluation factors, located in the diagonal of the square matrix, were calculated using the Spearman Correlation method. In order for convergent validity to be demonstrated, the diagonal values must be statistically significant to substantiate agreement between students and faculty. Inspection of Table 1 shows this not to be the case for all evaluation factors. Two of the seven evaluation factors were statistically significant at the p < .05 level with all others failing to meet this criterion. Group Interaction and Individual Rapport yielded significant correlations of .55 and .59, respectively.

Next, the existence of discriminant validity was assessed using two general guidelines outlined by Campbell and Fiske (1959). The first condition requires that each of the convergent validity coefficients be higher than the correlations between the different rating factors assessed by the two groups. An examination of the rows and columns of the square matrix will determine whether or not convergent validities are higher. Inspection of Table 3 reveals that the first
<table>
<thead>
<tr>
<th>FACTOR</th>
<th>ITEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Communication</td>
<td><strong>C/1</strong> - Instructor presents course material in an understandable manner.</td>
</tr>
<tr>
<td></td>
<td><strong>C/2</strong> - Instructor uses various visual and audio devices (e.g., chalkboard, overhead projector, etc.) to enhance the presentation of course material.</td>
</tr>
<tr>
<td></td>
<td><strong>C/3</strong> - Instructor speaks in a clear, concise manner.</td>
</tr>
<tr>
<td>Enthusiasm</td>
<td><strong>EN/4</strong> - Instructor conducts the course in an energetic, enthusiastic way.</td>
</tr>
<tr>
<td></td>
<td><strong>EN/5</strong> - Instructor enhances presentations with the use of humor.</td>
</tr>
<tr>
<td></td>
<td><strong>EN/6</strong> - Instructor's style of presentation holds your interest.</td>
</tr>
<tr>
<td>Organization/Planning</td>
<td><strong>O/7</strong> - Instructor returns exams and/or assignments within a reasonable period of time.</td>
</tr>
<tr>
<td></td>
<td><strong>O/8</strong> - Course materials are well prepared and carefully explained.</td>
</tr>
<tr>
<td></td>
<td><strong>O/9</strong> - The course material being taught follows the proposed objectives in a logical, sequential manner.</td>
</tr>
<tr>
<td>Group Interaction</td>
<td><strong>G/11</strong> - Instructor recognizes and acknowledges students' viewpoints.</td>
</tr>
<tr>
<td></td>
<td><strong>G/12</strong> - Students are encouraged to ask questions.</td>
</tr>
<tr>
<td></td>
<td><strong>G/13</strong> - Students are encouraged to participate in class discussions.</td>
</tr>
<tr>
<td>Individual Rapport</td>
<td><strong>I/14</strong> - Instructor responds to individuals in a courteous, friendly manner.</td>
</tr>
<tr>
<td></td>
<td><strong>I/15</strong> - Instructor is adequately accessible to students during office hours.</td>
</tr>
<tr>
<td></td>
<td><strong>I/16</strong> - Instructor encourages individuals to seek help/advice in or outside of class.</td>
</tr>
<tr>
<td>Subject Coverage</td>
<td><strong>S/17</strong> - Instructor presents the background or origin of ideas/concepts.</td>
</tr>
<tr>
<td></td>
<td><strong>S/18</strong> - Instructor answers questions in a meaningful, relevant manner.</td>
</tr>
<tr>
<td></td>
<td><strong>S/19</strong> - Instructor presents varying viewpoints (other than his/her own) when appropriate.</td>
</tr>
<tr>
<td>Examinations/Assignments</td>
<td><strong>E/20</strong> - Instructor provides feedback on exams and/or graded assignments</td>
</tr>
<tr>
<td></td>
<td><strong>E/21</strong> - Exams/graded materials tested course content as emphasized by the instructor.</td>
</tr>
<tr>
<td></td>
<td><strong>E/22</strong> - Methods of evaluating student work were fair and appropriate.</td>
</tr>
</tbody>
</table>

*Instructor self-evaluation forms were worded in first person.
TABLE 2

MEANS AND STANDARD DEVIATIONS OF INSTRUCTOR SELF- AND STUDENT EVALUATIONS

<table>
<thead>
<tr>
<th>FACTOR ITEM</th>
<th>INSTRUCTOR SELF-EVALUATIONS</th>
<th>STUDENT EVALUATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MEAN</td>
<td>STANDARD DEVIATION</td>
</tr>
<tr>
<td>Communication</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C/1</td>
<td>4.36</td>
<td>.63</td>
</tr>
<tr>
<td>C/2</td>
<td>3.64</td>
<td>.93</td>
</tr>
<tr>
<td>C/3</td>
<td>4.14</td>
<td>.66</td>
</tr>
<tr>
<td>Enthusiasm</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EN/4</td>
<td>4.30</td>
<td>.73</td>
</tr>
<tr>
<td>EN/5</td>
<td>4.00</td>
<td>.88</td>
</tr>
<tr>
<td>EN/6</td>
<td>4.00</td>
<td>.78</td>
</tr>
<tr>
<td>Org/Planning</td>
<td></td>
<td></td>
</tr>
<tr>
<td>O/7</td>
<td>4.62</td>
<td>.65</td>
</tr>
<tr>
<td>O/8</td>
<td>4.23</td>
<td>.60</td>
</tr>
<tr>
<td>O/9</td>
<td>4.46</td>
<td>.66</td>
</tr>
<tr>
<td>O/10</td>
<td>4.69</td>
<td>.48</td>
</tr>
<tr>
<td>Group Interaction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G-11</td>
<td>4.21</td>
<td>.70</td>
</tr>
<tr>
<td>G-12</td>
<td>4.14</td>
<td>.77</td>
</tr>
<tr>
<td>G-13</td>
<td>4.00</td>
<td>.96</td>
</tr>
<tr>
<td>Individual Rapport</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I-14</td>
<td>4.79</td>
<td>.43</td>
</tr>
<tr>
<td>I-15</td>
<td>4.14</td>
<td>.86</td>
</tr>
<tr>
<td>I-16</td>
<td>3.88</td>
<td>.95</td>
</tr>
<tr>
<td>Subject Coverage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S-17</td>
<td>3.71</td>
<td>.61</td>
</tr>
<tr>
<td>S-18</td>
<td>4.36</td>
<td>.63</td>
</tr>
<tr>
<td>S-19</td>
<td>4.14</td>
<td>.53</td>
</tr>
<tr>
<td>Examinations/Assignments</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E-20</td>
<td>4.62</td>
<td>.65</td>
</tr>
<tr>
<td>E-21</td>
<td>4.46</td>
<td>.66</td>
</tr>
<tr>
<td>E-22</td>
<td>4.70</td>
<td>.48</td>
</tr>
</tbody>
</table>
### TABLE 3

**CONVERGENT AND DISCRIMINANT VALIDITY OF STUDENT AND INSTRUCTOR EVALUATIONS FOR 14 COURSES**

<table>
<thead>
<tr>
<th>FACTORS</th>
<th>STUDENT EVALUATIONS</th>
<th>FACULTY SELF-EVALUATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3 4 5 6 7</td>
<td>8 9 10 11 12 13 14</td>
</tr>
<tr>
<td>Student Evaluations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Enthusiasm</td>
<td>84</td>
<td></td>
</tr>
<tr>
<td>2. GroupIn</td>
<td>.45</td>
<td>.80</td>
</tr>
<tr>
<td>3. Org/Pl</td>
<td>.65</td>
<td>.54</td>
</tr>
<tr>
<td>4. Indi Rap</td>
<td>.57</td>
<td>.58</td>
</tr>
<tr>
<td>5. Subco</td>
<td>.61</td>
<td>.61</td>
</tr>
<tr>
<td>6. Comm</td>
<td>.52</td>
<td>.46</td>
</tr>
<tr>
<td>7. Exam/Assign</td>
<td>.55</td>
<td>.54</td>
</tr>
<tr>
<td>Faculty Self-Evaluations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Enthusiasm</td>
<td>.38</td>
<td>.22</td>
</tr>
<tr>
<td>9. GroupIn</td>
<td>.19</td>
<td>.55*</td>
</tr>
<tr>
<td>10. Org/Pl</td>
<td>.57*</td>
<td>.19</td>
</tr>
<tr>
<td>11. Indi Rap</td>
<td>-.30</td>
<td>.007</td>
</tr>
<tr>
<td>12. Subco</td>
<td>-.25</td>
<td>-.41</td>
</tr>
<tr>
<td>13. Comm</td>
<td>.37</td>
<td>-.02</td>
</tr>
<tr>
<td>14. Exam/Assign</td>
<td>.20</td>
<td>-.01</td>
</tr>
</tbody>
</table>

*Statistically significant at p < .05
condition of discriminant validity has not been met as some of the convergent validities are lower than other correlations in the same row or column of the square matrix.

The second condition of discriminant validity requires that each convergent validity coefficient be higher than correlations between different traits assessed by the same group of raters. A comparison of the convergent validities and the off-diagonal correlations in the triangular matrices was made to determine the existence of the criterion. This condition was only partially met by both groups of raters. The data failed to find support for the second condition of discriminant validity which requires that each of the convergent validity coefficients be higher than correlations between different traits assessed by the same group.

In summary, the data failed to demonstrate the presence of convergent validity -- yielding but two statistically significant convergent validities. Secondly, the data assessed using two general guidelines outlined by Campbell and Fiske (1959) failed to provide evidence that student/faculty agreement on each factor was independent of agreement on other factors.
DISCUSSION

Evidence for convergent validity can be demonstrated if student and instructor ratings are substantially correlated. In this study, only two of the convergent validities (i.e., Group Interaction and Individual Rapport) were found to be statistically significant at the .05 level. Marsh (1982) reports that failure of the data to yield statistically significant correlations between student and self-evaluations, implies that the two different rating groups are measuring different constructs and a lack of validity exists in at least one of the methods.

The absence of discriminant validity in this study will be discussed in terms of the two criteria outlined earlier. The data failed to provide support for the first test of divergent validity which requires that each of the convergent validities be higher than any other correlation in the same row or column of the square matrix. As Marsh (1982) points out, failure to find such a condition could imply that agreement on a particular trait may be a generalized agreement on one or all of the traits. In this case, agreement would not have anything to do with the specific content of the rating factor itself. Instead, it could be a function of a generalized rating factor. The researcher concludes that agreement on one factor was not independent of agreement on other factors.

The results also failed to provide evidence of the second criterion of discriminant validity which requires the convergent validities to be
substantially higher than the correlations between different traits assessed by the same method. Failure of this test could indicate evidence of halo effect or method bias. As can be seen in Table 3, the intercorrelations among student evaluations or the off-diagonal correlations were rather high. These high intercorrelations could indicate that students were not discriminating among the various evaluation items. The instructor intercorrelations however, reflected substantially lower intercorrelations than that of students. This finding could indicate that instructors were less likely to evaluate themselves based on a generalized rating factor. Both Marsh (1982) and Marsh et al. (1979b) found evidence of some method or halo effect in student ratings of instruction.

Admittedly, there were several methodological problems which could have contributed to the inability of this study to demonstrate convergent and discriminant validity. The rating form contained separate groups of related items derived from a logical analysis of the content of effective teaching. It had not however, been previously factor analyzed to determine if the evaluation factors underlying the student evaluations were similar to those underlying faculty self-evaluations. Using an empirical procedure to demonstrate that the items within the same group do mean separate and distinguishable traits would have made interpretation possible. A well-defined factor structure might also have provided a safeguard against a halo effect. Future researchers would benefit from using such empirical procedures as factor analysis to determine that the instruments they have constructed provide measures of distinct components of teaching effectiveness. Where subject availability is a problem, as in this case, researchers might
want to consider using one of several instruments mentioned earlier (e.g., Marsh's SEEQ, Frey's Endeavor, etc.), each of which as a well-defined factor structure.

Other potential methodological problems in this study are within the instructor population. First, due to a restriction in the number of subjects available, the sample was relatively small and individuals who participated were not selected on a random basis. Those instructors who were asked to participate were primarily from the College of Arts and Sciences and more specifically from Psychology and Sociology departments. These individuals were also professors who were "most likely to participate". As can be seen in Table 2, means for instructor self-evaluations and student evaluations were for the most part on the high end of the rating scale (scale range is 1-5, with 1 lowest and 5 highest). These factors could have restricted the amount of variability found.

To what extent these methodological problems contributed in part, or in whole, to the failure of this study to demonstrate convergent and discriminant validity is questionable. The high intercorrelations found among student evaluations could also have had a significant impact on the outcome. Future researchers might want to consider training students how to rate prior to the actual evaluation of their instructors. Such training could reduce the amount of halo effect found.
November 17, 1986

Dear Human Subjects Committee:

The title of this thesis research project is Validity of Student Evaluations: A Comparison of Student Ratings to Instructor Self-Evaluations. This study will investigate the validity of students' evaluations of college instruction by comparing them to faculty self-evaluations using a multitrait-multimethod of analysis. The two groups of subjects will be University of Central Florida instructors and the students currently taking their courses. Students' evaluations and instructor self-evaluations will be collected during the Fall 1986 semester in undergraduate courses taught in the College of Arts & Sciences. University faculty will be asked to evaluate their own teaching proficiency with the same rating instrument completed by students. The evaluation instrument consists of 22 behavioral items and a five-point frequency scale will be used to rate each item (please see attached form). Evaluation forms will be disseminated and collected by the researcher or a designated student.

All subjects will be informed that data generated from this project will be used for a graduate thesis project and that this information will be used only by the researcher (myself) under the supervision of Dr. Wayne Burroughs. Subjects will also be assured that confidentiality will be maintained and individual identification will not be requested. A sign-up sheet will be available, at the time the survey is conducted, for those subjects who are interested in obtaining a summary of the results of this study.

Sincerely,

Becky Hopson
I/O Psy Graduate Student
Dear Student:

This is a thesis research project conducted by Becky Hopson under the supervision of Dr. Wayne Burroughs. The title of this project is Validity of Student Evaluations: A Comparison of Student Ratings to Instructor Self-Evaluations. This study will investigate the validity of students' evaluations of college instruction by comparing them to faculty self-evaluations. Students are asked to rate the instructor of the course they are currently taking, using the rating form distributed by the researcher.

The data generated from this project will be used by the researcher only. Confidentiality will be maintained and individual identification is not requested. Participation in this experiment is on a voluntary basis. If you choose to participate, please read carefully the attached Informed Consent Agreement and sign. A sign-up sheet will be made for those participants who are interested in obtaining a one-page summary of the results of this project. Your cooperation in this experiment will be greatly appreciated.

Sincerely,

Becky A. Hopson
I/O Pay Graduate Student
APPENDIX C

INFORMED CONSENT AGREEMENT

Directions: Please read the information below and sign your name in the space provided.

I hereby consent to participate in this research project and understand that the results of this evaluation will be used by Becky Hopson under the supervision of Dr. Wayne Burroughs. I have been informed that the data generated in this study will be used in an attempt to provide support for the validity of student ratings of college instruction. I have also been assured that confidentiality will be maintained and individual identification will not be requested. I understand that participation in this project is voluntary and that I may discontinue at any time.

Signature_____________________________________

Date:_________________________________________
INSTRUCTOR EVALUATION FORM

GENERAL INSTRUCTIONS

This evaluation form is intended to measure your observations of this instructors teaching proficiency. The purposes of obtaining this information are to provide data for a thesis research project and to facilitate the improvement of instructional quality by providing feedback to faculty members.

Read carefully each of the behavioral items and indicate the frequency with which you have observed him/her engage in that specific behavior. When you have finished, the researcher will collect the evaluation forms.

Using the rating scale below, circle the number that indicates the extent to which you believe the instructor has demonstrated this behavior.

<table>
<thead>
<tr>
<th>Rating Number</th>
<th>Performance Percentage</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0-64</td>
<td>Almost Never</td>
</tr>
<tr>
<td>2</td>
<td>65-74</td>
<td>Seldom</td>
</tr>
<tr>
<td>3</td>
<td>75-84</td>
<td>Sometimes</td>
</tr>
<tr>
<td>4</td>
<td>85-94</td>
<td>Frequently</td>
</tr>
<tr>
<td>5</td>
<td>95-100</td>
<td>Almost Always</td>
</tr>
</tbody>
</table>
(1) Instructor presents course materials in an understandable manner
   Almost Never  1  2  3  4  5  Almost Always

(2) Instructor uses various visual and audio devices (e.g., chalkboard, overhead projector, etc.) to enhance the presentation of course material
   Almost Never  1  2  3  4  5  Almost Always

(3) Instructor speaks in a clear, concise manner
   Almost Never  1  2  3  4  5  Almost Always

(4) Instructor conducts the course in an energetic, enthusiastic way
   Almost Never  1  2  3  4  5  Almost Always

(5) Instructor enhances presentations with the use of humor
   Almost Never  1  2  3  4  5  Almost Always

(6) Instructor's style of presentation holds your interest
   Almost Never  1  2  3  4  5  Almost Always

(7) Instructor returns exams and/or assignments within a reasonable period of time
   Almost Never  1  2  3  4  5  Almost Always

(8) Course materials are well prepared and carefully explained
   Almost Never  1  2  3  4  5  Almost Always

(9) The course material being taught follows the proposed objectives in a logical, sequential manner
   Almost Never  1  2  3  4  5  Almost Always

(10) The instructor meets his/her responsibility of holding class
    Almost Never  1  2  3  4  5  Almost Always

(11) Instructor recognizes and acknowledges students' viewpoints
    Almost Never  1  2  3  4  5  Almost Always
(12) Students are encouraged to ask questions  
Almost Never  1  2  3  4  5  Almost Always

(13) Students are encouraged to participate in class discussions  
Almost Never  1  2  3  4  5  Almost Always

(14) Instructor responds to individuals in a courteous, friendly manner  
Almost Never  1  2  3  4  5  Almost Always

(15) Instructor is adequately accessible to students during office hours  
Almost Never  1  2  3  4  5  Almost Always

(16) Instructor encourages individuals to seek help/advice in or outside of class  
Almost Never  1  2  3  4  5  Almost Always

(17) Instructor presents the background or origin of ideas/concepts  
Almost Never  1  2  3  4  5  Almost Always

(18) Instructor answers questions in a meaningful, relevant manner  
Almost Never  1  2  3  4  5  Almost Always

(19) Instructor presents varying viewpoints (other than his/her own) when appropriate  
Almost Never  1  2  3  4  5  Almost Always

(20) Instructor provides feedback on examinations and/or graded assignments  
Almost Never  1  2  3  4  5  Almost Always

(21) Examinations/graded materials tested course content as emphasized by the instructor  
Almost Never  1  2  3  4  5  Almost Always

(22) Methods of evaluating student work were fair and appropriate  
Almost Never  1  2  3  4  5  Almost Always
APPENDIX E

INSTRUCTOR SELF-EVALUATION FORM

GENERAL INSTRUCTIONS

This self-evaluation form is intended to measure the frequency in
which you feel you have demonstrated the following behaviors in this
course. The purposes of obtaining this information are to provide
data for a thesis research project and to facilitate the improvement
of instructional quality by providing feedback to faculty members.

Read carefully each of the behavioral items and indicate the
frequency with which you feel you have engaged in that specific
behavior. Do not report how you think students would rate you.
When you have finished, the researcher will collect the evaluation
forms.

Using the rating scale below, circle the number that indicates
the extent to which you believe you have demonstrated this behavior.

<table>
<thead>
<tr>
<th>Rating Number</th>
<th>Performance Percentage</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0–64</td>
<td>Almost Never</td>
</tr>
<tr>
<td>2</td>
<td>65–74</td>
<td>Seldom</td>
</tr>
<tr>
<td>3</td>
<td>75–84</td>
<td>Sometimes</td>
</tr>
<tr>
<td>4</td>
<td>85–94</td>
<td>Frequently</td>
</tr>
<tr>
<td>5</td>
<td>95–100</td>
<td>Almost Always</td>
</tr>
</tbody>
</table>
(1) I present course material in an understandable manner
Almost Never 1 2 3 4 5 Almost Always

(2) I use various visual and audio devices (e.g., chalkboard, overhead projector, etc.) to enhance the presentation of course material
Almost Never 1 2 3 4 5 Almost Always

(3) I speak in a clear, concise manner
Almost Never 1 2 3 4 5 Almost Always

(4) I conduct the course in an energetic, enthusiastic way
Almost Never 1 2 3 4 5 Almost Always

(5) I enhance presentations with the use of humor
Almost Never 1 2 3 4 5 Almost Always

(6) My style of presentation holds students' interest
Almost Never 1 2 3 4 5 Almost Always

(7) I return exams and/or assignments within a reasonable period of time
Almost Never 1 2 3 4 5 Almost Always

(8) My course materials are well prepared and I carefully explain them to students
Almost Never 1 2 3 4 5 Almost Always

(9) The course material I teach follows the proposed objectives in a logical, sequential manner
Almost Never 1 2 3 4 5 Almost Always

(10) I meet my responsibility of holding class
Almost Never 1 2 3 4 5 Almost Always

(11) I recognize and acknowledge students' viewpoints
Almost Never 1 2 3 4 5 Almost Always

(12) I encourage students to ask questions
Almost Never 1 2 3 4 5 Almost Always

(13) I encourage students to participate in class discussions
Almost Never 1 2 3 4 5 Almost Always
(14) I respond to individuals in a courteous, friendly manner
   Almost Never 1 2 3 4 5 Almost Always

(15) I am adequately accessible to students during office hours
   Almost Never 1 2 3 4 5 Almost Always

(16) I encourage individuals to seek help/advice in or outside of class
   Almost Never 1 2 3 4 5 Almost Always

(17) I present the background or origin of ideas/concepts
   Almost Never 1 2 3 4 5 Almost Always

(18) I answer students' questions in a meaningful, relevant manner
   Almost Never 1 2 3 4 5 Almost Always

(19) I present varying viewpoints (other than my own) when appropriate
   Almost Never 1 2 3 4 5 Almost Always

(20) I provide feedback on examinations and/or graded assignments
   Almost Never 1 2 3 4 5 Almost Always

(21) My examinations and graded materials tested course content as emphasized by the instructor
   Almost Never 1 2 3 4 5 Almost Always

(22) My methods of evaluating student work were fair and appropriate
   Almost Never 1 2 3 4 5 Almost Always
REFERENCES


