INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI
A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600
STUDENT EVALUATIONS OF OVERALL TEACHING EFFECTIVENESS: CAN SINGLE-ITEM MEASURES BE JUSTIFIED?

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in the Graduate School of The Ohio State University

By

Michael J. Hudy, B.A., M.A.

*****

The Ohio State University

1997

Dissertation Committee:
John P. Wanous, Adviser
James T. Austin
Robert S. Billings

Approved by
Advisor
Department of Psychology
ABSTRACT

In the field of industrial and organizational (I/O) psychology, the use of single-item measures to assess psychological constructs has been widely criticized based on the assumption that these measures are not psychometrically sound. Although multiple item scales are preferable to single-item measures in most situations, single-item measures may represent viable alternatives when one seeks to attain global unidimensional information about a construct. To support this contention, the validity and reliability of a single-item measure of overall teaching effectiveness were assessed.

The convergent validity of the single-item measure was assessed by correlating the single-item measure with a nine item scale of overall teaching effectiveness. Through use of the meta-analytic technique, mean weighted correlations between .78 and .80 were found. When corrected for the unreliability of the scale, estimates ranged from .82 to .84. The validity of the single-item measure was also
supported by the results of a factor analysis. The minimum reliability of the single-item measure was estimated using both the correction for attenuation formula and the results of a factor analysis. Estimates derived through use of the correction for attenuation formula ranged from .79 to .88, depending on the assumptions made. The single-item reliability estimate based on the results of a factor analysis was found to be .94.
Dedicated to Mom and Dad
ACKNOWLEDGMENTS

I cannot thank my adviser, John Wanous, enough for all of the support, encouragement, and invaluable feedback provided during this project. I also thank Jim Austin, Bob Billings, and Mary Roznowski for being members of my committee and for providing such useful feedback.

I thank Fred Ruland for all the valuable time he spent working with me setting up my data set. I also thank Michael Browne for his expert statistical consultation.

I am grateful to Richard Gunther for access to the data set used to conduct my empirical study.

I would also like to thank Joe Marth and Dennis Cue for their valuable insights.

Aaron, Monica, and Bo - thank you for your love and support.

And to Erin, thank you for your love and endless patience.
VITA

December 4, 1969............Born, Youngstown, Ohio

1992.........................B.A., Ohio Wesleyan University

1992-present.............Graduate Research Associate,
                          The Ohio State University

1993-1997...............Graduate Teaching Associate,
                          The Ohio State University

1995.........................M.A., The Ohio State University

PUBLICATIONS

Research Publications

1. J.P. Wanous, A.E. Reichers, and M.J. Hudy, "Overall job
   satisfaction: How good are single-item measures?" J. of

FIELDS OF STUDY

Major Field: Psychology
            Studies in Industrial/Organizational Psychology
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>ii</td>
</tr>
<tr>
<td>DEDICATION</td>
<td>iv</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>v</td>
</tr>
<tr>
<td>VITA</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vii</td>
</tr>
<tr>
<td>CHAPTERS:</td>
<td></td>
</tr>
<tr>
<td>1. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>Purpose of the Measure</td>
<td>7</td>
</tr>
<tr>
<td>Validity of the Measure</td>
<td>17</td>
</tr>
<tr>
<td>Reliability of the Measure</td>
<td>26</td>
</tr>
<tr>
<td>Practicality of the Measure</td>
<td>32</td>
</tr>
<tr>
<td>Present Study</td>
<td>35</td>
</tr>
<tr>
<td>2. METHOD</td>
<td>38</td>
</tr>
<tr>
<td>Sample</td>
<td>38</td>
</tr>
<tr>
<td>Instrument</td>
<td>39</td>
</tr>
<tr>
<td>Meta-Analysis</td>
<td>41</td>
</tr>
<tr>
<td>Factor Analysis</td>
<td>47</td>
</tr>
<tr>
<td>3. RESULTS</td>
<td>48</td>
</tr>
<tr>
<td>Descriptive Statistics</td>
<td>48</td>
</tr>
<tr>
<td>Validity Analyses</td>
<td>48</td>
</tr>
<tr>
<td>Meta-Analysis Results</td>
<td>48</td>
</tr>
<tr>
<td>Factor Analysis Results</td>
<td>55</td>
</tr>
<tr>
<td>Reliability Analyses</td>
<td>58</td>
</tr>
<tr>
<td>Use of the Correction for Attenuation Formula</td>
<td>58</td>
</tr>
<tr>
<td>Communaliti Estimates</td>
<td>59</td>
</tr>
<tr>
<td>4. DISCUSSION</td>
<td>60</td>
</tr>
<tr>
<td>Validity Results</td>
<td>60</td>
</tr>
<tr>
<td>Reliability Results</td>
<td>65</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Examples of multidimensional scales of teaching effectiveness</td>
</tr>
<tr>
<td>2</td>
<td>Individual level means, standard deviations, and item intercorrelations</td>
</tr>
<tr>
<td>3</td>
<td>Class mean level means, standard deviations, and item intercorrelations</td>
</tr>
<tr>
<td>4</td>
<td>Meta-analysis results</td>
</tr>
<tr>
<td>5</td>
<td>Moderator analysis using subset method</td>
</tr>
<tr>
<td>6</td>
<td>Moderator analysis using correlational method</td>
</tr>
<tr>
<td>7</td>
<td>Factor loading and communalities</td>
</tr>
</tbody>
</table>
CHAPTER 1

INTRODUCTION

The use of single-item measures to assess attitudes and beliefs has been widely criticized in the field of industrial and organizational (I/O) psychology (Sackett & Larson, 1990). The rationale used most often to support this criticism is that the internal consistency reliability of these measures cannot be determined and is probably unacceptably low. This rationale is based on the well-known theory of measurement error which holds that increasing the number of items on a scale will enhance reliability (Nunnally, 1978). Even though there is no empirical evidence showing that single-item measures are unreliable, these measures have been strongly discouraged in both research and practice. Although internal consistency reliability is an important consideration when choosing among measurement instruments, it should not be the only consideration. As Nunnally (1978, p. 192) states:

"For two reasons, it is easy to overstate the importance of the theory of measurement error in psychological measurement. First, measurement error
frequently does not harm most investigations as much as might be thought. Second, there are numerous topics regarding psychological measurement that are equally or more important than the theory of measurement error. A large portion of journal articles on psychological measurement and a major portion of some books on the topic have been devoted to measurement error. This is probably because the theory of measurement error is so neatly expressible in mathematical terms, in contrast to some other important issues, e.g., validity, where grounds for argument are not so straightforward. The theory of measurement error is one important topic in psychological measurement."

Single-item measures can be used as alternatives for two kinds of multiple-item scales: multidimensional and unidimensional. The major issue of concern differs for each of the two scales. When comparing a single-item measure to a multidimensional scale, the major issue is that of content validity. To justify the use of a single-item measure, it must be shown that the single-item measure can capture the domain of interest as fully as a multidimensional scale. When comparing a single-item measure to a unidimensional scale, the main concern becomes that of reliability. To justify the use of a single-item measure as an alternative to a unidimensional scale, the single-item measure must be shown to have reliability comparable to that of a unidimensional scale.

The purpose of the current study is to demonstrate that the unequivocal criticism of single-item measures may be overstated. The selection of an appropriate measure should
involve the consideration of the purpose, the validity, the reliability, and the practicality of the measure. The following literature review will establish that single-item measures may indeed be appropriate for certain purposes. When one desires unidimensional information about a particular construct, single-item measures may represent viable substitutes to both multidimensional and unidimensional scales. To support this contention, validity and reliability evidence will be presented. Finally, it will be argued that practical concerns sometimes justify the use of single-item measures.

The selection of a measure should be a process in which multiple pieces of information are weighed carefully. Information to be considered includes: (1) the purpose or intended use of the measure, (2) the validity of the measure, (3) the reliability of the measure, and (4) the practicality of the measure. Although the consideration of these four criteria is desirable, measurement tools are often selected on the basis of reliability and, to a lesser extent validity, with purpose and practicality receiving little, if any, consideration. By addressing each of the above four concerns, researchers and practitioners are able to best identify measures that will have the greatest value for their intended purposes.
Although the use of single-item measures is widely criticized in I/O psychology, few attempts have been made to either substantiate or refute such criticism. One exception is in the area of job satisfaction. Scarpello and Campbell (1983) argue that the use of a single-item measure of overall job satisfaction is acceptable by claiming that a single global measure of job satisfaction is more inclusive measure than the sum of specific job facets. They bolster their argument by claiming that empirical evidence of the unreliability of single-item measures of job satisfaction does not exist.

Wanous, Reichers, and Hudy (1997) extend this work by considering all four of the issues raised above: purpose, validity, reliability, and practicality. It is argued that global single-item measures of job satisfaction are better for purposes such as assessing change in overall job satisfaction over time. The use of specific facet scores may cause one to overlook changes in overall job satisfaction over time. For example, person A at time one has an average score of three by responding three to each of five questions. At time two, this same person may answer one to two questions, five to two questions, and three to one question, for an average of three. By using an average score across facets, it would be concluded that there has not been a change in overall job satisfaction over time.
Such a conclusion would be accurate if the changes in facet satisfactions "balanced themselves out" and had no effect on overall job satisfaction. It is more likely though that changes in facet satisfactions would affect overall job satisfaction. One way to directly address such a question would be to use a single-item measure of overall job satisfaction at both time one and time two.

The validity of single-item job satisfaction measures is demonstrated through a meta-analysis of correlations between single-item measures and various scales of job satisfaction. The scales of job satisfaction included multidimensional facet measures, sum of facet measures, and unidimensional overall measures of job satisfaction. The average uncorrected correlation was found to be .63 (SD = .09), .67 (SD = .08) when corrected for attenuation due to unreliability of the scale. Because the validity of job satisfaction scales is well-accepted, the strength of this correlation supports the validity of single-item measures of job satisfaction.

In addition to both purpose and validity considerations, Wanous et al. (1997) also address the reliability of single-item measures of job satisfaction. The minimum level of single-item reliability is estimated by using the correction for attenuation formula. These estimates range from .45 to .69, depending on the
assumptions made. Although these estimates are not high by conventional standards, they are probably much higher than critics would assume. Again, these are only minimum level estimates.

In regard to practicality, Wanous et al. (1997) cite several examples of situations in which single-item measures would be preferred over multiple-item scales. Limitations of questionnaire space and cost associated with items used in a telephone poll are two such examples. Taken as a whole, these results suggest that there may be situations in which single-item measures of overall job satisfaction can be justified as alternatives to unidimensional and multidimensional scales.

An area in the psychological literature where considerable debate has arisen over the use of single-item versus multidimensional scale measures is student evaluation of teachers. This debate has focused almost entirely on the use of student evaluations for personnel decisions such as tenure, promotion, and pay raises. The two sides of the issue can be summarized as follows. One camp argues that teaching is multifaceted and therefore must be evaluated using multidimensional scales (e.g. Marsh, 1987, 1991a, 1991b, 1994). The other camp counters that personnel decisions are unidimensional by nature and should be
measured as such. The most commonly used unidimensional measures of teaching effectiveness are global single-item measures. One major advantage of using these global single-item measures is that they are characterized by at least moderately high levels of validity across many different situations (e.g. Abrami, 1989; Cashin & Downey, 1992). The arguments posited by each camp will be more fully developed by examining each of the four measure-selection considerations mentioned earlier: purpose, validity, reliability, and practicality.

**Purpose of the Measure**

When selecting a measure of performance, the purpose of the evaluation is an essential consideration. This importance is reflected in the Principles for the Validation and Use of Personnel Selection Procedures (1980) developed by the American Psychological Association, Division of Industrial-Organizational Psychology. It is stated in these standards, "Criteria should be related to the purpose of the investigation" (American Psychological Association, 1980, p. 3). Although these principles were developed to address selection measures, they also apply to the measurement of performance (Landy & Farr, 1983). When performance measures
are used to make personnel decisions (e.g. promotion), they are, in effect, acting as selection measures.

Consideration of the purpose of a measure also will help determine whether a multidimensional or unidimensional approach is warranted. Some purposes require measures that allow one to gather detailed information about multiple dimensions of a construct. Job performance information collected for developmental purposes is an example of a situation in which one would be interested in collecting specific information about multiple dimensions of a construct. Single-item measures should not be used in situations such as these because they cannot provide the detailed information desired. Other purposes are better served by unidimensional measures. An overall rating of job performance to be used for a promotion decision is an example of a situation that would call for a unidimensional measure. When unidimensional information is sought, single-item measures may indeed be an acceptable alternative to multi-item scales.

The evaluation of instruction may serve one of three primary purposes: (1) to help the instructor improve, (2) to provide information to colleagues and administrators for personnel decisions, and (3) to provide information to students for course selection (Brandenburg, Braskamp, & Ory, 1979). More generally, the purpose of instructional
evaluation can be viewed as either summative or formative. Summative evaluation involves the use of student ratings to make a final judgment about the instructor's teaching effectiveness. Personnel decisions such as promotion, tenure, and salary increases involve summative evaluation. Formative evaluation is diagnostic in that instructors use these ratings to assess areas of possible improvement (Cashin & Downey, 1992).

When selecting a student evaluation measure, the purpose for which the data are going to be used should guide the choice of the measure. This means that certain measures may be preferable for formative evaluation while other measures may be preferable for summative evaluation. For formative evaluation, there is general agreement that an appropriately constructed multidimensional scale is more useful than global unidimensional measures (Marsh, 1991b, 1994). This is because such scales provide specific information on multiple factors that are believed to contribute to teaching effectiveness. These scales allow instructors to pinpoint potential areas of weakness that can be targeted for improvement. Some examples of the dimensions represented by these multidimensional scales are presented in Table 1.

As for summative evaluation, there is not as much agreement as to which type of measure is most effective.
<table>
<thead>
<tr>
<th>Instrument</th>
<th>Dimension Names</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frey’s Endeavor instrument</td>
<td>Presentation Clarity, Workload, Personal Attention, Class Discussion, Organization/Planning, Grading, Student Accomplishments</td>
</tr>
<tr>
<td>The Student Description of Teaching (SDT) questionnaire</td>
<td>Analytic/Synthetic Approach, Organization/Clarity, Instructor Group Interaction, Instructor Individual Interaction, Dynamism/Enthusiasm</td>
</tr>
<tr>
<td>Marsh’s SEEQ instrument</td>
<td>Learning/Value, Instructor Enthusiasm, Organization, Individual Rapport, Group Interaction, Breadth of Coverage, Examinations/Grading, Assignments/Readings, Workload/Difficulty</td>
</tr>
<tr>
<td>The Michigan State SIRS instrument</td>
<td>Instructor Involvement, Student Interest &amp; Performance, Student-Instructor Interaction, Course Demands, Course Organization</td>
</tr>
</tbody>
</table>

Note:
Adapted from Marsh (1987)

Table 1: Examples of Multidimensional Scales of Teaching Effectiveness
Marsh (1987) argues that because teaching is multifaceted, teaching effectiveness should not be summarized by ratings on a global unidimensional measure. According to this argument, a unidimensional measure of instructor effectiveness cannot fully capture the multifaceted domain of teaching. Summative evaluation should instead be made on the basis of ratings received on specific factors included on a multidimensional scale. By using a multidimensional scale to assess instruction, it is believed that the domain of effective teaching is more fully represented.

Others such as Abrami (1985, 1989), Cashin and Downey (1992), and Scriven (1981) argue that summative evaluation is best served by unidimensional measures such as global single-item measures. According to this view, summative evaluation is nothing more than a simple request for an overall judgment of an instructor's effectiveness, so it should be measured as such. Cashin and Downey (1992) suggest that single-item measures of teaching effectiveness are similar to final course grades for students in that both are aimed at providing a usable measure of the level of overall performance. Because student effectiveness is multidimensional and can still be summarized by one final course grade, it makes sense to evaluate instructor effectiveness in much the same way. This idea is supported by Schmidt and Kaplan (1971) in that they suggest
performance of an individual can be summarized by a single number for use in making personnel decisions.

Abrami (1985) states his case for the use of single-item measures of instructor effectiveness by calling to attention that summative decisions about teaching are almost always unidimensional, not multidimensional. The example he uses is that of a tenure committee that is presented with information that a faculty member has varying degrees of organization, concern, knowledge, enthusiasm, and stimulation. The tenure committee still must summarize all of this information into a single measure of teaching effectiveness. Even though teaching is multidimensional, summative decisions about teaching effectiveness usually are not.

This reasoning can be extended to a literature more familiar to I/O psychologists, that of assessment centers. The assessment center method is commonly used to evaluate management ability and potential. During the course of an assessment center, candidates are evaluated on a variety of dimensions (e.g. oral and written communication, leadership, planning, etc.), but also typically receive an Overall Assessment Rating (Huck & Bray, 1976). Although managerial performance is multidimensional, this multidimensional information must be combined to arrive at one unidimensional summative hiring decision.
The above reasoning has been interpreted by Marsh (1991a) as questioning the multidimensional nature of teaching effectiveness. This is not the case. Abrami (1985) recognizes that teaching is indeed multidimensional, but suggests that ratings of teaching can be either multidimensional or unidimensional, depending on how these ratings will be used. If ratings are to be used to diagnose specific areas of strength or weakness of a particular instructor, then a multidimensional approach is definitely warranted. If, on the other hand, ratings are to be used to make a personnel decision based on overall teaching effectiveness, then a unidimensional conceptualization may be preferable.

Abrami (1985) uses Wechsler's (1958) approach to the measurement of intelligence as an analogy to the measurement of teaching effectiveness: Wechsler believed that intelligence consists of a combination of specific components that form one's general intellectual ability. Teaching effectiveness can be viewed in much the same manner; multiple components that combine to form overall teaching effectiveness. Extending this analogy, some intelligence tests are designed to assess specific components of intelligence, such as verbal ability, spatial ability, or mathematical ability. Other intelligence tests
are designed to provide a measure of overall intelligence, such as IQ tests.

This is not to say that because intelligence can be viewed as unidimensional, that it can be accurately assessed with a single-item measure. Intelligence is much too complex for such measurement. Most would argue that teaching effectiveness is not nearly as complex as intelligence. Sackett and Larson (1990) identify complexity as a key issue when it comes to single-item measurement in that the more complex and multidimensional a construct, single items become more difficult to interpret. Intelligence is a construct that is multifaceted and extremely complex and thus should not be measured with a single item. Conversely, teaching effectiveness is not as complex a construct and can be conceptualized as unidimensional for purposes such as summative evaluation.

Wanous et al. (1997) argue that there is a middle ground when it comes to the issue of construct complexity. Some constructs are so narrow and unambiguous that single-item measurement is widely accepted. Examples of such constructs would be intent to leave a job (Sackett & Larson, 1990) and expectancy theory components such as the belief that effort leads to performance (e.g. Ilgen, Nebeker, & Pritchard, 1981). Intelligence represents a construct at the opposite end of the complexity continuum in that this is
too complex for single-item measurement. Wanous et al. (1997) suggest that overall job satisfaction is a construct that falls somewhere in between these two extremes of the complexity continuum and thus single-item measurement may be acceptable. Teaching effectiveness would seem to be another construct that lies in this middle ground.

One final rationale for the use of unidimensional measures of teaching effectiveness for summative evaluation is that there is a lack of sophisticated theories linking multidimensional scale items to effective teaching (Abrami, 1985, 1989). Instead, scale items are often selected by faculty and student committees which use their own implicit theories of effective teaching to generate item pools. Scales developed in this manner may result in factors that are unrelated to effective teaching or factors that are only related to effective instruction in specific contexts. For example, course difficulty items are often found on student evaluation forms, but this factor has been shown to have no relationship with criteria of teaching effectiveness (Cohen, 1981). An example of a factor that may only predict effective instruction in specific situations is rapport (Abrami, 1985, 1989; Cohen, 1981). Rapport may contribute to effective teaching when class sizes are small, but it is difficult to establish rapport when there are 300 students in a class. Making summative judgments on the basis of
factors such as these is problematic and should be avoided. An instructor who consistently teaches very large classes would not be pleased to discover that a personnel decision such as tenure was based, in part, on ratings of the rapport established with the class. Another related problem associated with including items unrelated to effective teaching is that these items may bias the responses to other items (Scriven, 1981). If students provide low ratings for items of factors unrelated to teacher effectiveness, this may have an impact on the ratings provided for items of factors that actually do predict effective performance. Until theories of effective course instruction are developed to guide item selection, it may be appropriate to base summative evaluations on unidimensional measures, such as global single-items.

In summary, when choosing a measure of teaching effectiveness, the purpose of the evaluation is one important consideration. For formative evaluation, multidimensional scales are preferable measures. When the purpose is summative evaluation, it may be better to use a unidimensional measure of teaching effectiveness, such as a global single-item measure; especially until theories of teaching effectiveness are developed and tested.
Validity of the Measure

A second important consideration when selecting a teaching effectiveness measure is that of validity. Abrami, d'Apollonia, and Cohen (1990) suggest that the validity of student ratings can be approached in two ways. First, student ratings can be viewed as measures of student satisfaction with instruction. In this view, measures are valid if they accurately represent student satisfaction with quality of instruction. In the second approach, student ratings are viewed as measures of performance and are considered valid only if they accurately reflect teaching effectiveness.

There are very few validity studies of single-item student ratings as measures of student satisfaction, because ratings are infrequently used in this manner. When viewed as measures of student satisfaction with instruction, student ratings are similar to ratings of job satisfaction. The validity of single-item measures of job satisfaction has been supported by Scarpello and Campbell (1983) and Wanous et al. (1997).

Scarpello and Campbell (1983) suggest that single-item measures of overall job satisfaction actually may be more valid than satisfaction scores attained by summing the facets of multidimensional scales (e.g. Job Description
Index). This is because multidimensional scales may not be content valid when it comes to the measurement of overall job satisfaction. What this means is that the measurement of overall job satisfaction with a multidimensional scale is limited by the facets assessed by the scale (e.g. promotion, supervisor, pay, etc.). These facets may not be inclusive of all the factors that influence a person's level of overall job satisfaction. Similarly, these scales may include facets that do not affect overall job satisfaction. Measuring overall job satisfaction with a single global item allows one to avoid these potential content validity problems because no attempt is made to identify factors that contribute to overall job satisfaction. This reasoning can be applied to ratings of overall student satisfaction with instruction. Single-item measures may be more valid because they are not limited by factors believed to contribute to overall satisfaction with instruction.

Building upon the work of Scarpello and Campbell (1983), Wanous et al. (1997) assessed the convergent validity of single-item measures of overall job satisfaction. This was done by conducting a meta-analysis of job satisfaction studies reporting both scale and single-item measures of overall job satisfaction. Scale measures included both unidimensional scales and multidimensional scales. The mean observed correlation between scales and
single-item measures was found to be .63, .67 when corrected for scale unreliability. This correlation was found to be highest (.72) when single-item measures were compared to unidimensional scales of job satisfaction. The strength of this correlation indicates that single-item measures of overall job satisfaction may be an acceptable alternative to multi-item scales, especially when such scales are unidimensional. Although this type of study is yet to be done in the domain of student evaluations of teaching, it is expected that the results will generalize.

Although student ratings of instruction can be thought of as measures of satisfaction with instruction, these ratings are most often viewed as measures of teaching effectiveness. Because of this, many validity studies have been conducted attempting to link student ratings to numerous teaching effectiveness criteria. Although there is disagreement among researchers, student achievement is generally considered the most important product of teaching effectiveness (Cohen, 1981). The most common measures used as an index of student achievement are grade on a final course exam and final course grade. These measures are not perfect indexes of student achievement because they are influenced by factors that cannot be controlled by an instructor; factors such as student ability and motivation (Cohen, 1981). Because of this, one would expect validity
coefficients of the relationship between student ratings and student achievement to be modest.

Prior to the 1980's, validity studies of student ratings of instructor effectiveness were summarized using narrative reviews. These reviews resulted in inconsistent findings that could not be used to make overall conclusions. A better technique to synthesize the results of psychological research is that of meta-analysis (Hunter & Schmidt, 1990). In contrast to the qualitative voting method used in narrative reviews, meta-analysis is a quantitative procedure that allows one to calculate an overall observed effect size for an area of research characterized by many primary studies. Additionally, meta-analysis allows one estimate the true effect size of a research domain by correcting for study imperfections, or artifacts.

Cohen (1981) conducted a meta-analysis of validity studies examining the relationship between student ratings of instruction and student achievement. The meta-analysis was conducted only with studies that used a multisection design to assess the validity of student ratings of instruction. The multisection design uses courses that have multiple section offerings so that setting effects such as syllabus, textbook, and section size are minimized. The measure of student achievement used in multisection designs
is that of grade on a final examination common to all sections. This minimizes potential bias due to instructor grading standards. One final strength of the multisection design is that class sections are used as the unit of analysis, rather than students. This level of analysis allows one to assess the effect of the instructor on student ratings and student achievement (Abrami et al., 1990). A total of 41 validity studies employing the multisection design were used in the Cohen (1981) meta-analysis. These 41 independent studies reported relationships between student ratings and student achievement for 68 multisection courses.

Cohen (1981) reports results for two global, unidimensional ratings (overall course and overall instructor) and seven specific rating dimensions (Skill, Rapport, Structure, Difficulty, Interaction, Feedback, and Evaluation). The two overall ratings were most commonly assessed using global single-item measures. The average correlations between the two overall ratings and student achievement were found to be among the highest (.43 for overall instructor and .47 for overall course). The magnitude of these average correlations is moderately high, given that there are many factors outside the control of the instructor that influence student achievement. Only one specific dimension rating, Skill (.50), was found to have a
higher average correlation with student achievement. Five of the seven specific dimension rating correlations were .31 or lower and all five of these included zero in the 95% confidence interval. These results suggest that global unidimensional ratings, which are often assessed using single-item measures, may be the most valid measures of teaching effectiveness.

One explanation of these results is that overall ratings are more generalizable than specific rating factors (Abrami, 1985, 1989a, 1989b; Abrami & d'Apollonia, 1991; Cashin, Downey, & Sixbury, 1994). An overall rating would be expected to be a valid measure of teaching effectiveness across a wide variety of courses, instructors, students, and settings. The same cannot be said about specific rating factors. The use of specific factors as teaching effectiveness measures implies that there is one strongly supported teaching method that is effective across all settings and situations (Cashin et al., 1994). This simply is not the case. For example, specific factors such as Rapport and Interaction may indeed be valid measures of instructor effectiveness in a small class setting, but are unlikely to be valid measures in a large class setting. When examining the validity of specific rating factors, it may be more appropriate to establish the local validity of
the factor (Abrami, d'Apollonia, & Cohen, 1990). That is, it may be more accurate to speak of the validity of specific rating factors only under certain conditions. If specific rating factors are to be used as measures of teaching effectiveness, their validity must be determined in each setting and situation in which they are going to be used. Contrary to this, the validity of overall ratings is likely to generalize across a wide variety of settings and situations.

The results of the Cohen (1981) meta-analysis can also be explained by examining the specificity of the measures used in the study. According to Fisher (1980), the strongest relationships between predictors and criteria will be observed when the specificity of their measurement matches. Applying this to the validity of teaching effectiveness measures, global measures should best predict global criteria, while specific measures should best predict specific criteria. In the Cohen (1981) study, the criterion was an overall measure of student achievement. It should not be surprising then that the global measures of teaching effectiveness demonstrated stronger relationships with this criterion than specific measures.

In accordance with these meta-analysis results, Abrami (1985) has suggested that global single-item ratings may account for a substantial portion of the variance in
measures of teaching effectiveness. This assertion was tested by Cashin and Downey (1992) in a study of 17,183 classes from 105 institutions. Regression analyses were used to estimate the percentage of variance in a self-reported measure of student progress that could be accounted for by global single-item ratings and specific factor ratings. Separate regression analyses were run for each of two global single-item ratings (overall course and overall instructor). The results show that each of the two single-item measures were able to account for at least 50% of the variance in the criterion variable: course, 60% \((r = .77)\) and instructor, 54% \((r = .74)\). Although these global single-item measures were able to account for a substantial portion of the variance in the criterion, adding specific factors led to an increase (14% and 9% respectively for instructor and course) in the explained variance.

Additionally, the Cashin and Downey (1992) study is somewhat limited by the use of student ratings for both predictor and criterion variables. This study does, however, provide preliminary support for Abrami's (1985) suggestion that global single-item measures can account for a great deal of variance in teaching effectiveness criteria.

Although the two studies described above provide support for the validity of single-item measures of teaching
effectiveness, it has been suggested that these measures are valid only because they often appear after students have responded to a series of specific factor items (Marsh, 1994). The reasoning here is that specific factor items create a frame of reference for students that allow them to make more accurate overall ratings. Similarly, it is suggested that global single-item measures of teaching effectiveness will not be valid if they are used apart from specific factor items. Ory (1982) empirically tested this relationship between global single-item measures and specific factor items by placing the single-item measure either before or after many specific items. The results of this study indicated that there are no significant differences in global single-item measures due to placement. This implies that students have a set opinion of an instructor's effectiveness that does not require a frame of reference created by specific factor items (Ory, 1982). This finding provides support for the use of single-item measures in isolation of specific factor items.

In summary, evidence supports the validity of single-item measures of teaching effectiveness. Single-item measures have been found to have moderately high validity coefficients and to explain a significant portion of variance in criterion measures. Global single-item measures used together with some specific factor items explain a
greater portion of variance in criterion measures, but global single-item measures alone account for over 50% of the variance. It has also been demonstrated that global single-item measures are consistent whether they are used as part of a multi-item scale or used in isolation.

Reliability of the Measure

Estimates of reliability can be grouped into three general classes: (1) measures of stability, (2) measures of equivalence, and (3) measures of internal consistency (Landy & Farr, 1983). The most common measure of stability is that of retest reliability. Although retest reliability can be calculated for single-item measures, researchers are not overly concerned with retest reliability because the appropriate magnitude depends on the nature of the construct being measured (Wanous & Reichers, 1996). This measure is not given as much attention as internal consistency reliability because acceptable levels of retest reliability are arguable whereas internal consistency reliability has well-established acceptable levels. However, this does not mean that retest reliability should be dismissed. Retest reliability can still be used as an indicator of the quality of a measure. A high level of retest reliability of a single-item measure could be used as one piece of evidence to support the psychometric soundness of such a measure.
However, retest reliability levels for single-item measures of teaching effectiveness are rarely reported.

The second class of reliability estimates is that of measures of equivalence. One commonly used measure of equivalence that is applicable to single-item measures of teaching effectiveness is interrater reliability. According to Remmers (1931), reliability of student ratings can be thought of as the level of agreement among different students rating the same instructor. This means that increasing the number of students providing ratings of an instructor will lead to an increase in interrater reliability. Remmers (1931) argues that the Spearman-Brown equation can be applied to the majority of situations in which subjective judgments are used to determine the number of judgments required to attain a given level of reliability. In the case of single-item measures of teaching effectiveness, the Spearman-Brown equation could be applied to determine the number of students required to meet a predetermined acceptable level of reliability.

The main argument against the use of single-item measures is that it is not possible to estimate the internal consistency reliability (Sackett & Larson, 1990; Wanous & Reichers, 1996; Wanous et al., 1997). Wanous et al. (1997) present a technique for estimating the reliability of single-item measures. This can be done by using the well
known formula for the correction for attenuation in a novel way. This formula is expressed as:

\[ r_{xy}^\hat{} = \frac{r_{xy}}{\sqrt{r_{xx} * r_{yy}}} \]

where, \( r_{xy} \) = correlation between variables \( x \) and \( y \), \( r_{xx} \) = reliability of variable \( x \), \( r_{yy} \) = reliability of variable \( y \), and \( r_{xy}^\hat{} \) = the assumed "true" underlying correlation between variable \( x \) and \( y \) if both were measured perfectly. This formula is most commonly applied when variable \( x \) and variable \( y \) represent two distinct constructs for the purpose of estimating how much the correlation would be if the two variables were completely reliable. However, the formula also applies when two variables are both measuring the same construct, such as a single-item and a multiple-item measure of teaching effectiveness. In this case, Nunnally (1978, p. 220) said that, "The correlation between two such tests would be expected to equal the product of the terms in the denominator and consequently \( r_{xy}^\hat{} \) would equal 1.00...If \( r_{xy}^\hat{} \) were 1.00, \( r_{xy} \) would be limited only by the reliabilities of the two tests: \( r_{xy} = \sqrt{r_{xx} * r_{yy}}. \)"

The observed correlation between single-item and scale measures of teaching effectiveness can be put into this formula, along with the observed reliability of the scale measure. By assuming the underlying correlation of the two
measures of teaching effectiveness is 1.00, one can estimate the reliability of the single-item measure by solving the equation for the missing value (reliability of the single-item measure). Applying the formula in this manner would result an estimate of the minimum reliability of the single-item measure because it is based on the assumption that the underlying correlation between the single-item and scale measures of teaching effectiveness is 1.0 (Wanous & Reichers, 1996; Wanous et al., 1997).

Estimates of single-item reliability have been calculated in this manner in two studies. First, Wanous and Reichers (1996) used the correction for attenuation formula to estimate the reliability of three single-item measures: overall job satisfaction, perceived amount of participation in decision-making, and desired amount of participation in decision-making. Reliability estimates of these three single-item measures were also calculated at two times, for a total of six single-item reliability estimates. The average minimum reliability for these six measures was found to be .57. When the assumption of perfect reliability between the single-item and scale measure was relaxed to .90, the average reliability estimate for the single-item measures increased to .70.

In the second study, conducted by Wanous et al. (1997), reliability estimates of single-item measures of overall job
satisfaction were calculated. The data used for this study came from the results of a meta-analysis of overall job satisfaction conducted by the authors. From the meta-analysis, the mean uncorrected correlation between single-item and scale measures of overall job satisfaction, along with the observed reliabilities of the scale measures were used in the correction for attenuation formula. This resulted in estimates of single-item reliability ranging from .45 to .69, depending on the assumptions made.

Taken together, the results of the two studies described above are promising. Although the reliability of single-item measures overall job satisfaction and participation in decision-making may not be as high as well-constructed scales, these estimates represent only minimum reliabilities and are probably higher than many researchers would assume. Depending on the assumptions made, some of the reliabilities approach levels that are considered acceptable by many researchers. Although the estimation of single-item reliabilities in this manner is promising, such estimates are yet to be calculated for single-item measures of teaching effectiveness.

Another way that the reliability of a single-item measure can be estimated involves the use of factor analysis. In factor analysis, the total variance of a
variable is made up of the communality, the specificity, and the unreliability (Harman, 1967). The communality is the variance in a measured variable accounted for by common factors, the specificity is the amount of variance in a measured variable accounted for by unique factors, and the unreliability is the amount of variance in a measured variable accounted for by error of measurement. The reliability of a variable, then, is the communality plus the specificity. Based on this, Harman (1967, p. 19) states, "In other words, the communality of any variable is less than or equal to the reliability of the variable, and equals the reliability only when the specificity vanishes."

According to Weiss (1976), there are three types of communality estimates: (1) reliability estimates, (2) highest correlation of a variable with others, and (3) the squared multiple correlation. Reliability estimates are the largest of the three because they include both specific and common variance. The highest correlation of a variable with other variables estimates is based on the assumption that a variable cannot correlate with any other variable higher than its own reliability. This estimate represents a more conservative reliability coefficient because a variable usually correlates with other variables lower than its own reliability. The squared multiple correlation estimate is the most conservative in that it is the squared multiple
correlation of a variable with all other variables in the matrix. This description reiterates that the communality of a variable can be used as a conservative estimate of the reliability of that variable. This is yet another unexplored avenue in the reliability estimation of single-item measures of teaching effectiveness.

In summary, three types of reliability information can be attained for single-item measures of teaching effectiveness. Retest reliability can be calculated for single-item measures, but is not done so often because acceptable levels are not established and vary from construct to construct. According to Remmers (1931), interrater reliability of single-item measures of instructor effectiveness can be enhanced by increasing the number of students providing ratings. Finally, it has been shown that the internal consistency reliability of a single-item measure can be estimated in two ways: (1) by using the correction for attenuation formula and (2) by using communality estimates derived from factor analysis. These procedures are yet to be applied to the study of single-item measures of teaching effectiveness.

Practicality of the Measure

One final piece of information to consider when deciding upon an appropriate measure is that of
practicality. Although the technical and psychometric aspects of a measure are important, one must also consider practical issues such as cost, usability, and credibility. There are numerous instances in both research and practice when practical limitations favor the use of single-item measures over lengthy scales.

In regard to cost, the use of a single-item measure will almost always compare favorably to the use of a scale. Examples of cost savings achieved with use of single-item measures include lower copying costs, less data entry time, and savings of time and money associated with shorter telephone polls. The first two of these examples are especially relevant to student evaluations of instructors. Abrami et al. (1990) note that the length of forms used to assess teaching effectiveness range from a low of one item to a high of 141 items. The cost difference between the use of a single-item measure and a 141-item scale would be tremendous in terms of length of printing, collating, keypunching or scanning, synthesizing, reporting, and interpreting (Scriven, 1981). In this sense, the selection of a measure can be viewed as a decision of return on investment. The researcher or practitioner must consider whether the extra information gained by using lengthy scales is worth the increased cost.
A second practical concern when selecting a measure is usability. The ability of the end consumer to use and understand the information gathered by a measure should factor into the selection of such a measure. Teaching effectiveness measures are often used by university administrators to make personnel decisions. Such administrators who are often nonexperts in evaluation cannot be expected to make personnel decisions on the basis of assigning weights to factor scores attained from multidimensional scales of teaching effectiveness (Abrami, 1989a, 1989b). Administrators may also prefer the simplicity of interpreting a single overall score over interpreting a score on a unidimensional scale. A single-item measure represents a much more understandable and usable measure for nonexperts in evaluation.

One final practical concern is the credibility or perceived fairness of the measure. This issue refers to the face validity of a measure. In some cases, single-item measures may be perceived as more face valid than scales consisting of multiple items. In the evaluation of teaching effectiveness, the issue of face validity can be viewed from both the instructor and student point of view. From the instructor's standpoint, being rated on a single overall teaching effectiveness item will likely be perceived as fair because such ratings are simple and straightforward. From a
student perspective, being asked to provide an overall rating of the instructor may seem more justified than having to respond to a series of items that appear to be repetitious. Responding to multiple items that are viewed as repetitious can lead to irritation and attempts at hypothesis guessing. In certain contexts, single-item measures offer greater face validity than multiple-item scales.

Present Study

The present study represents a follow-up and extension of the Wanous et al. (1997) study of single-item measures of overall job satisfaction. It is proposed that, similar to overall job satisfaction, overall teaching effectiveness is a construct that can be adequately measured with a single global item. The basis for this argument is that when used for summative evaluation, overall teaching effectiveness can be conceived as a unidimensional construct of moderate complexity, a level of complexity not great enough to preclude it from single-item measurement. To support this contention, the validity and reliability of a single-item measure of overall teaching effectiveness will be assessed.

The validity of the single-item measure of teaching effectiveness will be assessed in two ways. First, the single-item measure will be correlated with a unidimensional
scale of teaching effectiveness. Because such multiple-item scales are generally accepted to be valid measures of teaching effectiveness, it would follow that a strong correlation between a single-item measure and a unidimensional scale would support the validity of the single-item measure.

Second, a factor analysis of the ten items (nine scale and one global) will be conducted. This will be done to examine the pattern of factor loadings on an overall teaching effectiveness factor. It is believed that the global single-item measure will load heavily on this factor, more heavily than any of the nine scale items. In other words, an underlying overall teaching effectiveness factor should influence the global single-item measure more so than the scale items. This should be true because the scale items are influenced by both this general teaching effectiveness factor as well as other unique specific factors. On the other hand, most of the variance of the global single-item should be explained by the overall teaching effectiveness factor.

The reliability of the single-item measure of teaching effectiveness will be estimated with two different techniques. First, as described above, the correction for attenuation formula will be adapted to calculate the minimum reliability of the single-item measure. Second, from the
results of the factor analysis, the communality of the single-item measure can be interpreted as a conservative estimate of its reliability.
CHAPTER 2

METHOD

Sample

Student ratings ($N = 323,262$ students) were collected during 1994 and 1995 at the Ohio State University. These ratings were collected at the main Columbus campus, as well as five branch campuses. Students included undergraduates, graduates, and professional students. Classes were taught by instructors at a variety of levels ranging from graduate teaching associates to full professors. Because ratings were entered into a data set at the individual student level only, these ratings had to be aggregated to form classes. To form the classes, ratings were aggregated on the basis of seven variables that were coded identically for each member of a particular class. If individuals were found to be equal on all seven of these variables, they were grouped into a class. These variables were class size, course level, average grade received in the class, ethnicity of the instructor, sex of the instructor, campus, and the rank of the instructor. These variables are more fully described
below. Classes from which fewer than five ratings were obtained were excluded, as were classes with missing data. Complete data were available for a total of 10,682 classes.

After forming the classes, a measure of convergence was calculated to assess the accuracy of the aggregation process. The estimator of within-class agreement used was that of $r_{wg}$ which is an index of within group agreement for a set of judges rating a single target with a single item (James, Demaree, & Wolf, 1984, 1993). This index compares the amount of variation in the observed response pattern to the amount of variation expected if responses were completely random. When using the average level of within-class variation found in the current study (.19), $r_{wg}$ was found to be .90. The magnitude of this index provides support for quality of the aggregation procedure used to form the class-level ratings.

**Instrument**

The Student Evaluation of Instruction (SEI) form was used to obtain the student ratings. The SEI was designed to be a unidimensional scale of teaching effectiveness to be used explicitly for the purpose of collecting data appropriate for use in making personnel decisions such as promotion, tenure, and merit-pay decisions. These teaching
effectiveness ratings were to be used as part of a broader performance assessment program that would include various forms of peer evaluation (Gunther, 1996). The instrument was developed through the joint effort of five different committees comprised of 43 faculty, staff, and student members. The development process included face-to-face interviews with faculty and administrators, telephone interviews with faculty and students, an exhaustive review of the teaching evaluation literature, and two pretests of the instrument.

This process resulted in a 10-item evaluation form that included nine scale items and one global item, "Overall, I would rate this instructor as..." The nine scale items represent facets of overall teaching effectiveness and were evaluated on a five-point Likert scale ranging from Disagree Strongly (1) to Agree Strongly (5). For these scale items, a Not Applicable response area was also provided. The global single-item was evaluated on a five-point Likert scale ranging from Poor (1) to Excellent (5). In addition to these ten items, the SEI asks students for their class standing (Freshman, Sophomore, etc.), their cumulative grade point average, and their reason for enrolling in the class (required, free elective, etc.). A copy of the SEI is included in the Appendix.
In addition to the SEI responses provided by students, descriptive information about the instructor and the course was included in the data set. This was accomplished through an interface with the university's central records kept in the registrar's database. When SEI forms were generated for use by each instructor, descriptive information about the instructor and the course was automatically coded onto the evaluation forms. This was done through the use of a call number that was unique for each course taught by each instructor.

Meta-Analysis

To demonstrate the validity of the single-item measure of teaching effectiveness, the meta-analytic technique was used to calculate the average correlation across classes between the 9-item scale and the single-item measure of instructor effectiveness. By using meta-analysis, the observed average correlation could be corrected for artifacts such as unreliability and potential moderators of the relationship could be identified.

All meta-analyses were conducted by utilizing the META 5.3 software developed by Schwarzer (1989). This software uses the Hunter and Schmidt (1990) method of meta-analysis and can be used to combine probabilities, $d$ or $g$ effect
sizes, and correlational effect sizes. The accuracy of the program has been carefully tested by reproducing results published in the meta-analysis literature. One limitation of this particular software is that it cannot analyze more than 500 effect sizes. Because of this limitation, random samples of 500 classes were drawn from the entire sample of classes. The characteristics of the subsamples were compared to the characteristics of the entire sample of classes and found to be practically identical. Based on this, it is believed that results derived from a random sample of 500 classes will generalize to the entire sample of classes.

Separate meta-analyses were conducted for each of five different sample size cutoffs. First, a meta-analysis was run on a random sample of 500 classes selected from the entire sample (N = 5 or greater). Because this analysis included many small sample sizes, it was decided that the meta-analysis would be run again with sample size minimums of 20, 40, 60, and 80. With cutoffs of 20 and 40, random samples of 500 classes were again selected. When the cutoff was raised to a sample size of 60, 227 classes remained and were used in the meta-analysis. Similarly, when the cutoff was raised to a sample size of 80, 92 classes remained and were used for the meta-analysis.
To conduct a meta-analysis on correlational effect sizes, one needs the following information: the sample size, the observed correlation between variables X and Y, the observed reliability of variable X, and the observed reliability of variable Y. If reliability information is not available for one of the variables, the most conservative practice is to substitute 1.0 as the average reliability for this variable. In regard to the reliability of the single-item measure of teaching effectiveness, this conservative approach was used in one set of meta-analyses. In a second set of meta-analyses, single-item reliabilities were estimated using the correction for attenuation formula, as described above. With the five different sample size cutoffs and two approaches to single-item reliability, a total of ten meta-analyses were run.

In addition to an examination of the main effects, a moderator analysis was also conducted. This was accomplished by coding class characteristics that had the potential of affecting the relationship between the two measures of teaching effectiveness. The following variables were coded for the moderator analysis:

1. **Class Size.** This variable was determined by the official university enrollment numbers for each class.
Class sizes ranged from five to 491, with an average of 28.4 (SD = 23.2).

2. Course Level. This refers to the numeric value assigned to each class to represent its academic level (e.g. 100's represent introductory-level classes).

3. Campus. Ratings were collected at the main Columbus campus and five branch campuses. For the purpose of the moderator analysis, this variable was dichotomized into either "Main Campus" or "Branch Campus." Ratings for 9,026 (84.5%) classes at the main campus and 1,656 (15.5%) at the branch campuses were used in the analysis.

4. Instructor Rank. Ratings were provided for five levels of experience. The ratings were scaled as follows: professors (5), associate professors (4), assistant professors (3), lecturers (2), and graduate teaching associates (1). Ratings were attained for 1,293 (12.1%) classes taught by professors, 2,090 (19.6%) by associate professors, 1,453 (13.6%) by assistant professors, 1,448 (13.6%) by lecturers, and 4,398 (41.2%) by graduate teaching associates.

5. Ethnicity of the Instructor. Because the great majority of instructors were Caucasian White, this variable was dichotomized into two categories: "White" or "Not White". The second category included Black, Native
American, Asian, and Hispanic instructors. A total of 8,442 (79.0%) classes were taught by an instructor of "White" ethnicity, while 2,240 (21.0%) classes were taught by an instructor of "Not White" ethnicity.

6. Average Grade of Class. This variable was attained by averaging the numeric grade values (e.g. A = 4.0) assigned to each student of a particular class. Average Grade of Class ranged from 0.75 to 4.0, with an average of 2.96 (SD = 0.62).

A search for moderator variables can be accomplished in one of two ways. First, class characteristics can be used to separate the data into subsets with separate meta-analyses run on each of these subsets. Moderator variables are characterized by large differences in mean effect sizes between subsets, and a reduction in variance within subsets (Hunter & Schmidt, 1990). The second method is to correlate the class characteristic with effect size. This correlation is systematically reduced by sampling error and can be corrected by dividing the observed correlation between class characteristic and effect size by the square root of the effect size reliability. The effect size reliability is attained by dividing the effect size variance in the population by the observed effect size variance (Hunter & Schmidt, 1990).
Both of these methods were used in the present study to search for potential moderator variables. The subset method was used when class characteristic variables could be divided into meaningful categories. The potential moderator variables of Campus and Ethnicity of the Instructor were analyzed using the subset method. The Class Size, Course Level, Instructor Rank, and Average Grade of Class variables could not be grouped into two or three categories without a significant loss of information, so these potential moderator variables were tested using the correlational method described above.

All moderator analyses were conducted using the sample of classes with a class size cutoff of five. This sample was chosen because it is characterized by the least amount of range restriction in the potential moderating variables. As the class size cutoff points are raised, the Class Size, Course Level, Instructor Rank, and Campus variables become restricted to specific levels. For example, large classes are much more common at low course levels than at high course levels. To accurately represent the full range of the potential moderator variables, the sample with a class size cutoff of five was used.
Factor Analysis

A factor analysis of the nine scale items and the global single-item was conducted using class mean ratings as the basis of the analysis. According to Cranton and Smith (1990), variations in ratings at the class mean level reflect perceived differences in teaching effectiveness among instructors. When individual student ratings within a class are used as the unit of analysis, variations in ratings reflect individual differences in the perceptions of students. By using class mean ratings, these individual differences among students' perceptions are eliminated. SPSS for Windows (Release 6.1) was used to conduct the factor analysis on a total of 10,682 classes. Principle axis factoring was used as the method of extraction.
CHAPTER 3

RESULTS

Descriptive Statistics

Descriptive statistics for the nine-item scale and the single global item are presented in Tables 2 and 3. Table 2 includes means, standard deviations, and item intercorrelations at the individual student level (N = 239,509). All correlations are significant at the p<.01 level and range from .48 to .80. Table 3 includes means, standard deviations, and item intercorrelations at the class mean level (N = 10,682). All correlations are significant at the p<.01 level and range from .55 to .93.

Validity Analyses

Meta-Analysis Results

The results of the ten meta-analyses are presented in Table 4. The upper half of the table includes the results of the five meta-analyses conducted with 1.0 as the single-item reliability estimate. The lower half of the table includes the results of the five meta-analyses conducted
<table>
<thead>
<tr>
<th>Item</th>
<th>X</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organized</td>
<td>4.19</td>
<td>.95</td>
</tr>
<tr>
<td>Stimulating</td>
<td>3.94</td>
<td>1.07</td>
</tr>
<tr>
<td>Interested</td>
<td>4.41</td>
<td>.86</td>
</tr>
<tr>
<td>Think</td>
<td>4.18</td>
<td>.94</td>
</tr>
<tr>
<td>Prepared</td>
<td>4.33</td>
<td>.89</td>
</tr>
<tr>
<td>Helping</td>
<td>4.27</td>
<td>.96</td>
</tr>
<tr>
<td>Learned</td>
<td>3.95</td>
<td>1.09</td>
</tr>
<tr>
<td>Atmosphere</td>
<td>4.07</td>
<td>1.01</td>
</tr>
<tr>
<td>Communicated</td>
<td>4.06</td>
<td>1.09</td>
</tr>
<tr>
<td>Overall</td>
<td>4.15</td>
<td>.96</td>
</tr>
</tbody>
</table>

Note:
N = 239,509
All correlations are significant at the p<.01 level

Table 2: Individual level means, standard deviations, and item intercorrelations
<table>
<thead>
<tr>
<th>Item</th>
<th>X</th>
<th>SD</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organized</td>
<td>3.98</td>
<td>.55</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stimulating</td>
<td>4.00</td>
<td>.58</td>
<td>.66</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interested</td>
<td>4.09</td>
<td>.62</td>
<td>.73</td>
<td>.67</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Think</td>
<td>4.12</td>
<td>.53</td>
<td>.61</td>
<td>.70</td>
<td>.76</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prepared</td>
<td>4.20</td>
<td>.54</td>
<td>.85</td>
<td>.59</td>
<td>.79</td>
<td>.65</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Helping</td>
<td>4.21</td>
<td>.49</td>
<td>.67</td>
<td>.55</td>
<td>.86</td>
<td>.72</td>
<td>.72</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Learned</td>
<td>4.25</td>
<td>.45</td>
<td>.80</td>
<td>.80</td>
<td>.84</td>
<td>.77</td>
<td>.80</td>
<td>.80</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Atmosphere</td>
<td>4.34</td>
<td>.49</td>
<td>.80</td>
<td>.70</td>
<td>.86</td>
<td>.79</td>
<td>.81</td>
<td>.85</td>
<td>.90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Communicated</td>
<td>4.36</td>
<td>.48</td>
<td>.81</td>
<td>.64</td>
<td>.80</td>
<td>.71</td>
<td>.83</td>
<td>.78</td>
<td>.87</td>
<td>.89</td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>4.45</td>
<td>.44</td>
<td>.81</td>
<td>.70</td>
<td>.87</td>
<td>.78</td>
<td>.85</td>
<td>.85</td>
<td>.92</td>
<td>.93</td>
<td>.92</td>
</tr>
</tbody>
</table>

Note:
N = 10,682
All correlations are significant at the p < .01 level

Table 3: Class mean level means, standard deviations, and item intercorrelations
Table 4: Meta-analysis results

<table>
<thead>
<tr>
<th>Sample Classes</th>
<th>N=5</th>
<th>N=20</th>
<th>N=40</th>
<th>N=60</th>
<th>N=80</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Classes</td>
<td>10,682</td>
<td>3,668</td>
<td>681</td>
<td>227</td>
<td>92</td>
</tr>
<tr>
<td>Total N</td>
<td>9,546</td>
<td>16,220</td>
<td>30,473</td>
<td>20,021</td>
<td>10,901</td>
</tr>
<tr>
<td>Weighted Mean r</td>
<td>.79</td>
<td>.78</td>
<td>.79</td>
<td>.80</td>
<td>.80</td>
</tr>
<tr>
<td>SD</td>
<td>.16</td>
<td>.16</td>
<td>.12</td>
<td>.12</td>
<td>.12</td>
</tr>
<tr>
<td>95% Confidence Interval</td>
<td>.53-1.05</td>
<td>.49-1.06</td>
<td>.57-1.02</td>
<td>.59-1.02</td>
<td>.57-1.03</td>
</tr>
<tr>
<td>Corrected r</td>
<td>.84</td>
<td>.82</td>
<td>.83</td>
<td>.84</td>
<td>.84</td>
</tr>
<tr>
<td>SD</td>
<td>.14</td>
<td>.15</td>
<td>.12</td>
<td>.11</td>
<td>.12</td>
</tr>
<tr>
<td>95% Confidence Interval</td>
<td>.57-1.11</td>
<td>.53-1.12</td>
<td>.60-1.06</td>
<td>.62-1.07</td>
<td>.61-1.08</td>
</tr>
<tr>
<td>% of Variance Due to</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sampling Error</td>
<td>29.86</td>
<td>18.36</td>
<td>14.89</td>
<td>10.61</td>
<td>7.34</td>
</tr>
<tr>
<td>Unreliability</td>
<td>5.29</td>
<td>4.21</td>
<td>2.18</td>
<td>2.44</td>
<td>3.80</td>
</tr>
<tr>
<td>Total</td>
<td>35.14</td>
<td>22.57</td>
<td>17.07</td>
<td>13.05</td>
<td>11.14</td>
</tr>
</tbody>
</table>

With Single-Item
Reliability Estimates

| Corrected r    | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| SD             | .00 | .06 | .05 | .05 | .05 |
| 95% Confidence Interval | 1.0-1.0 | .88-1.12 | .90-1.10 | .90-1.10 | .89-1.11 |
| % of Variance Due to | | | | | |
| Sampling Error  | 29.86| 18.36| 14.89| 10.61| 7.34 |
| Unreliability   | 85.14| 72.79| 73.67| 76.67| 79.65 |
| Total           | 115.00| 91.15| 88.57| 87.28| 86.99 |
with single-item reliabilities estimated using the
correction for attenuation formula.

Using 1.0 as the reliability estimate for the single-
item measure, meta-analyses were conducted with five
different class size cutoffs. When a class size cutoff of
five or greater is used, the mean observed correlation
between the scale and single-item is .79 (SD = .16). When
this is corrected for unreliability of the scale, the
correlation is .84 (SD = .14). The amount of between-class
variance due to all artifacts is 35%, of which 30% is
attributed to sampling error. Changing the class size
cutoffs to 20, 40, 60, and 80 has little effect on the
results of the meta-analysis. With these higher cutoffs,
the mean observed correlations range from .78 to .80, with
standard deviations ranging from .12 to .15. When corrected
for unreliability of the scale, correlations range from .82
to .84 and standard deviations range from .11 to .15. As
the class size cutoffs are raised, the percentage of
between-class variance explained by artifacts goes down, due
in most part to a reduction in sampling error.

When the single-item reliability is estimated using the
correction for attenuation formula, the corrected
correlations change significantly. For all five of the
class size cutoffs, the correlation corrected for
unreliability of both the scale and the single-item is 1.0
(standard deviations range from 0 to .06). The total amount of between-class variance explained by artifacts range from 87% to over 100%. This large increase in the percentage of between-class variance explained by artifacts is due entirely to the increase in the percentage of between-class variance attributed to unreliability.

The second phase of the meta-analysis consisted of a search for potential moderators. The subset method was used to test two potential moderator variables, Ethnicity of Instructor and Campus. The results of this analysis are presented in Table 5. For Ethnicity of the Instructor, two subsets were formed, “White” and “Not White”. The observed correlation is .79 (SD = .16) for the “White” subsample and .80 (SD = .17) for the “Not White” subsample. Two subsets were also formed based on the Campus variable, “Main Campus” and “Branch Campus”. The observed correlation is .79 (SD = .15) for the “Main Campus” subset and .76 (SD = .21) for the “Branch Campus” subset. When using the subset method, a moderator is indicated by subset differences in the mean observed correlations and a reduction of the between-class variance. Because this is not the case in both Ethnicity of Instructor and Campus subset analyses, it is concluded that these variables do not represent moderators of the relationship between single-item and scale measures of overall teaching effectiveness.

53
<table>
<thead>
<tr>
<th>Ethnicity of Instructor</th>
<th>Classes</th>
<th>Total N</th>
<th>Weighted Mean r</th>
<th>SD</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>406</td>
<td>7,918</td>
<td>.79</td>
<td>.16</td>
<td>.53-1.04</td>
</tr>
<tr>
<td>Not White</td>
<td>94</td>
<td>1,628</td>
<td>.80</td>
<td>.17</td>
<td>.50-1.10</td>
</tr>
<tr>
<td>Main</td>
<td>432</td>
<td>8,469</td>
<td>.79</td>
<td>.15</td>
<td>.55-1.04</td>
</tr>
<tr>
<td>Branch</td>
<td>68</td>
<td>1,077</td>
<td>.76</td>
<td>.21</td>
<td>.41-1.12</td>
</tr>
</tbody>
</table>

Table 5: Moderator analysis using subset method
Four potential moderator variables were tested using the correlational method: Class Size, Course Level, Instructor Rank, and Average Grade of Class. The results of this analysis are presented in Table 6. The observed correlations range from -.07 to .04, -.08 to .04 when corrected for sampling error. Because of the small magnitude of these correlations, it is concluded that none of these four variables moderate the relationship between single-item and scale measures of overall teaching effectiveness.

Factor Analysis Results

The results of the factor analysis conducted with the ten overall teaching effectiveness items (nine scale and one global) are presented in Table 7. All ten items load heavily on one Overall Teaching Effectiveness factor. Of particular interest for the purpose of demonstrating the validity of the single-item measure is the factor loading of the global item compared to the average loadings of nine scale items. The single global item has the highest loading at .97, compared to an average loading of .87 for the nine scale items.
Table 6: Moderator analysis using correlational method

<table>
<thead>
<tr>
<th></th>
<th>X</th>
<th>SD</th>
<th>Correlation w/ Effect Size</th>
<th>Corrected Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instructor Rank</td>
<td>3.53</td>
<td>1.49</td>
<td>-.06</td>
<td>-.06</td>
</tr>
<tr>
<td>Course Level</td>
<td>3.12</td>
<td>2.42</td>
<td>.04</td>
<td>.04</td>
</tr>
<tr>
<td>Average Grade</td>
<td>2.91</td>
<td>.65</td>
<td>-.07</td>
<td>-.08</td>
</tr>
<tr>
<td>Class Size</td>
<td>28.55</td>
<td>23.98</td>
<td>.01</td>
<td>.01</td>
</tr>
<tr>
<td>Item</td>
<td>Factor Loading</td>
<td>Communality</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>---------------</td>
<td>-------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organized</td>
<td>.85</td>
<td>.72</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stimulating</td>
<td>.74</td>
<td>.55</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interested</td>
<td>.90</td>
<td>.82</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Think</td>
<td>.81</td>
<td>.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prepared</td>
<td>.87</td>
<td>.75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Helping</td>
<td>.86</td>
<td>.73</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Learned</td>
<td>.95</td>
<td>.90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Atmosphere</td>
<td>.96</td>
<td>.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Communicated</td>
<td>.92</td>
<td>.84</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scale Average</td>
<td>.87</td>
<td>.76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall Item</td>
<td>.97</td>
<td>.94</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Analysis conducted at class-mean level

Table 7: Factor loadings and communalities
Reliability Analyses

Use of the Correction for Attenuation Formula

The reliability of the single-item measure of teaching effectiveness was estimated using the correction for attenuation formula, as described above. Because the results of the five meta-analyses using different class size cutoffs were so similar, it was determined that the single-item reliability estimate would be calculated using the results attained in the class size cutoff of five meta-analysis only. Repeating these calculations with the results from the other four meta-analyses would yield nearly identical results.

When using the correction for attenuation formula to estimate single-item reliability, the assumed underlying correlation between the scale and single-item measure must be entered into the formula. The most conservative estimate of this underlying construct correlation is to assume a perfect 1.0 correlation between the two measures. By using this most conservative estimate one assumes that if both measures were completely reliable, they would be correlated at a 1.0 level.

After assuming a construct correlation of 1.0, the other unknowns can be attained from the results of the meta-analysis. The average reliability of the nine item scale is .88. The observed weighted average correlation between the
scale and single-item measure is .79. The observed correlation is used rather than the corrected correlation because use of the corrected correlation would result in an inflated reliability estimate due to double-correction (Wanous et al., 1997). Using these numbers in the correction for attenuation formula results in an estimate of .71 as the minimum reliability of the single-item measure. If the assumed underlying construct correlation of 1.0 is relaxed somewhat to .90, then the minimum single-item reliability estimate increases to .88.

**Communality Estimates**

As described above, communality estimates derived from factor analysis can be used as minimum estimates of single-item reliability. The results of a factor analysis of the ten teaching effectiveness items (nine scale and one global) are presented in Table 7. The communality estimates range from .55 to .94. The average communality estimate for the nine scale items is .76, while the communality estimate for the single global item is .94.
The research presented here was conducted in an effort to demonstrate single-item measures can be used as an acceptable alternatives to commonly used scales. This is appropriate when the construct of interest is unidimensional rather than multidimensional. The construct of interest in the present study is that of overall teaching effectiveness. It is argued that for unidimensional summative purposes, single-item measures are acceptable. For practical reasons, these measures may even be preferred over lengthy scales. The results presented here provide strong support for the validity and reliability of single-item measures of overall teaching effectiveness.

Validity Results
To demonstrate the validity of a single-item measure of overall teaching effectiveness, the correlation between a single-item measure and a unidimensional scale was assessed. This was accomplished using a meta-analytic technique. This
resulted in an observed correlation ranging from .78 to .80, depending on the class size cutoff used. The magnitude of these correlations provides strong support for the convergent validity the single-item measure of teaching effectiveness.

When correlations were corrected for the unreliability of the scale, validity estimates ranged from .82 to .84, depending on the class size cutoff used. When these correlations were corrected for the unreliability of the scale and the single-item, the estimated correlation became 1.0 for all class size cutoffs. One explanation for this finding is as follows. For these analyses, single-item reliability was estimated using the correction for attenuation formula. As described earlier, this procedure yields a minimum estimate of single-item reliability. Viewed another way, this estimate is actually the maximum level of unreliability of the single-item measure. When applying the meta-analytic technique, the use of this conservative estimate of single-item reliability will result in the maximum amount of correction for unreliability. Because of this, the correlation between scale and single-item measures of overall teaching effectiveness may be overcorrected. When using single-item reliability estimates
for meta-analysis purposes, it may be best to use fairly liberal estimates to avoid potential overcorrection of effect sizes.

The analyses conducted to search for potential moderators of the relationship between single-item and scale measures of overall teaching effectiveness did not yield any significant results. For this particular analysis, the sample of classes with class size cutoff of five was used. In this sample, 35% of the between-class variance was due to sampling error and unreliability of the scale. This leaves 65% of the between-class variance unexplained. This remaining between-class variance can be explained in one of two ways. First, there may be unmeasured moderator variables that can account for this unexplained variance. Examples of such variables could include instructor teaching styles, student learning styles, and instructor grading standards. A second explanation for the remaining between-class variance is that the unreliability of the single-item measure accounts for a large portion of this unexplained variance. When estimates of single-item reliability derived from the correction for attenuation formula are included in the meta-analysis, nearly all of the between-class variance is accounted for (87% to over 100%, depending on class size cutoff). The most plausible explanation for the unaccounted between-class variance is that both of the above factors
explain the remaining between-class variance.

The second method used to demonstrate the validity of the single-item measure of overall teaching effectiveness involved the interpretation of the results of a factor analysis conducted with the ten items (nine scale and one global). Of particular interest was the pattern of factor loadings. The global single-item measure was found to have a factor loading of .97, compared to an average loading of .87 for the nine scale items. These factor loadings can be interpreted as beta weights that represent the linear influence of common factors on each of the measured variables (Nunnally, 1978). The scale used in the present study is unidimensional, so there is only one common factor that influences each of the ten items. Because this scale was designed to assess overall teaching effectiveness, there is good reason to believe that this common factor does indeed represent an underlying overall teaching effectiveness construct.

Based on this, one would expect an overall teaching effectiveness factor to influence a global measure of teaching effectiveness more so than an item that measures a facet of teaching effectiveness. This should be true because the items that make up the scale will surely be influenced by the overall teaching effectiveness factor, but these items are likely to also be affected by specific
factors other than overall teaching effectiveness. Because the single-item measure is global in nature, specificity is removed causing the overall teaching effectiveness factor to have a greater influence. Stated another way, the global single-item is a "cleaner" measure of the underlying teaching effectiveness construct than the items that comprise the scale. This probably would not be the case if the scale items were all global in nature (rather than facets). If this were the case, the pattern of factor loadings would be likely to be uniformly high.

When factor loadings are squared, they become equivalent to communality estimates and can be interpreted as the percentage of variance in the measured variables accounted for by the common factors (Nunnally, 1978). The communality of the global single-item is .94, compared to an average communality of .76 for the scale items. This means that 94% of the variance in the global single-item can be explained by the overall teaching effectiveness factor. This leaves only 6% of the variance that can be attributed to unique (specific) factors and measurement error. On the other hand, 76% of the variance in the scale items is explained by the overall teaching effectiveness factor. Although this is still a large portion of the overall variance, much more variance (24%) remains to be accounted for by specificity and error. This interpretation further
supports the proposition that the global single-item is a "cleaner" measure of overall teaching effectiveness than is the scale.

Taken together, the results of the meta-analysis and the factor analysis provide strong support for the validity of a single-item measure of overall teaching effectiveness. The large correlation between the single-item and unidimensional scale measures indicates convergent validity. The factor loadings and communality estimates show that the single-item measure is influenced almost entirely by the underlying overall teaching effectiveness factor. This indicates that the global single-item is a valid measure of overall teaching effectiveness.

Reliability Results

The reliability of a single-item measure of overall teaching effectiveness was estimated two ways. First, the correction for attenuation formula was used to calculate single-item reliability. Based on different assumptions, reliability estimates of .71 and .88 were derived. The lower of the two estimates was based on the most conservative assumption that the underlying construct correlation between the two measures of overall teaching effectiveness is 1.0. When this assumption is relaxed to an underlying construct correlation of .90, the reliability
estimate increased to .88.

The second method used to estimate single-item reliability involved the use of factor analysis. In factor analysis, the variance of a variable consists of the communality, the specificity, and the unreliability (Harman, 1967). Because of this, communality estimates can be used as a conservative estimate of a variable's reliability. A factor analysis of the ten (nine scale and one global) items yielded a communality estimate for the global single-item of .94. The average communality estimate for the nine scale items was found to be .76. The reason that the estimate for the global single-item is higher than the estimate for the scale items can be attributed to specificity. As mentioned earlier, the global single-item is less likely to be affected by specific factors because of its level of generality. The nine scale items are likely to be affected more by specific factors, which would lead to lower communality estimates.

Using the two different methods of single-item reliability estimation yields results that are somewhat different. There are a number of explanations for such a finding. First, the reliability estimates derived using the correction for attenuation formula may be low because the assumed construct correlations of 1.0 and .90 are both too conservative. The underlying construct correlation actually
may be lower than .90. By using .87 as the underlying construct correlation, one would attain the same reliability estimate (.94) that was found using factor analysis. Although this is possible, it is not very likely because a reliability of .94 would mean that the single-item is more reliable measure than the scale (.88).

Instead, it is more likely that the reliability estimate attained through factor analysis has been artificially inflated. One potential cause for such inflation would be the fact that the class means used to calculate the correlations used for the factor analysis are not weighted for sample size. When calculating the intercorrelations on which the factor analysis is based, each class mean is treated as one observation. Weighting each of the means for sample size could lower the magnitude of the correlations used for the factor analysis. Such a change in the pattern of intercorrelations would be likely to lower the communality estimates derived from the factor analysis.

One biasing factor that may act to inflate both of the single-item reliability estimates is that of halo effect. When rating teaching effectiveness, students may have a general idea of the overall effectiveness of the instructor and use this general impression to assign ratings to all items on an evaluation form, leading to consistent ratings.
across all items. This may lead to inflations in both of
the single-item reliability estimates, but would inflate
estimates derived by factor analysis to a greater extent.
This would be true for the following reason. Halo effect
would artificially inflate intercorrelations among facets,
the correlation between the scale and single item, and the
scale reliability. Single-item reliability estimates
calculated using the correction for attenuation formula rely
on the correlation between the scale and single-item measure
and the reliability of the scale. Increasing the magnitude
of the correlation between the scale and the single-item
would cause an increase in the single-item reliability
estimate, whereas increasing the reliability of the scale
would cause a decrease in the single-item reliability
estimate. In this case, the influence of inflation due to
halo effect would be minimal. On the other hand, single-
item reliability estimates derived from factor analysis are
dependent on the inter-item correlations. Artificially
inflating these inter-item correlations would lead to an
increase in the communality estimates that are used as
proxies for single-item reliability. Thus, single-item
reliability estimates derived from factor analysis are more
susceptible to the influence of halo effect. Although this
is the case, it is not possible to determine how much effect
that halo effect has on the item intercorrelations.

68
Integration with Previous Research

Integrating the results of the present study with the results of Wanous et al. (1997) provides strong support for the use of single-item measures. Single-item measurement may be justified when the purpose of a research project is to collect unidimensional information about a construct. The consideration of practical issues may further support the decision to use a single-item measure. Empirical evidence from the current research and the Wanous et al. (1997) study provides strong support for the validity and reliability of single-item measures.

In the present research, a single-item measure of overall teaching effectiveness was found to correlate strongly (.78 to .84) with a unidimensional scale of the same construct. A high factor loading (.97) of the single-item measure provided further validity evidence. Two methods of single-item reliability estimation yielded favorable results. Use of the correction for attenuation formula resulted in a reliability estimates ranging from .79 to .88, while factor analysis results yielded an estimate of .94. Wanous et al. (1997) found a correlation of .63 between single-item and scale measures of overall job satisfaction and estimated the single-item reliability to range from .45 to .69 by using the correction for attenuation formula.
Taken together, these results support the validity and reliability of single-item measures. When comparing the results of the two studies, it is apparent that the validity and reliability estimates found in the present study are significantly higher than similar estimates presented in the Wanous et al. (1997) research. One possible explanation for this difference is that the two studies focus on different constructs. This would mean that the single-item measure of overall teaching effectiveness simply is a more valid and reliable tool than the single-item measure of overall job satisfaction. Although this may be the case, the discrepancy between the two measures is not likely to be so large. The two measures are comparable in terms of complexity and both provide unidimensional information about their respective constructs.

A more plausible explanation for the validity and reliability discrepancy found in the two studies concerns the scale measures to which the single-item measures are compared. In the present study, a single-item measure of overall teaching effectiveness was compared to a unidimensional scale of the same construct. In the Wanous et al. (1997) study, single-item measures of overall job satisfaction were compared to three different types of scales: multifaceted scales, scales measuring one specific facet (Work facet), and global unidimensional scales. When
focusing only on the Wanous et al. (1997) results attained using global unidimensional scales as the comparison, findings are slightly more similar to those presented in the current study. The observed correlation of single-item measures of overall job satisfaction with these scales was found to be .67 (compared to .63 for the entire sample) and single-item reliability estimates ranged from .51 to .69 (compared to .45 to .61 for the entire sample). These results must be interpreted with caution because they are based on only eight effect sizes.

Comparing a global single-item measure to a global unidimensional scale provides the most informative and interpretable test of the single-item measure. This is because the purpose of both measures is to collect general, overall information about the construct of interest. When a single-item measure is compared to a multidimensional scale, the results are not easily interpreted. If one finds that there is a low correlation between the single-item and the multidimensional scale, this could mean that the single-item measure is not a valid measure of the underlying construct. On the other hand, this low correlation could be interpreted as the multidimensional scale being a poor measure of a global construct. This is the same point made by Scarpello and Campbell (1983). When a multifaceted scale is used to gather overall information about a construct, two problems
may arise. First, the scale may contain facets that are not related to the global construct. Second, the scale may leave out facets that are related to the global construct. Because of this, comparisons of single-item measures with unidimensional scales are more meaningful.

Limitations

There are several limitations to the research presented here. First, the generalizability of the results is limited because all data were collected at one university. The generalizability of the study is enhanced somewhat because student ratings were collected at the main campus as well as several branch campuses. Another generalizability concern stems from the fact that the single-item measure of overall teaching effectiveness was compared to only one specific unidimensional evaluation form. Findings could differ when single-item measures are compared to other unidimensional teaching effectiveness scales. Because of these potential limitations, this type of study should be replicated in different research settings using other unidimensional overall teaching effectiveness scales.

As mentioned earlier, another potential limitation of the current study is the possibility that halo effect has biased the student ratings of overall teaching effectiveness. An indication that such a bias may exist
comes from the pattern of moderate to strong intercorrelations observed among the nine scale items (see Tables 2 and 3). As described earlier, the presence of halo effect could potentially inflate the single-item validity and reliability estimates presented here.

Range restriction is another potential limitation of the present study. At the class mean level, the means of the ten items used to assess overall teaching effectiveness range from 3.98 to 4.45, with the global single-item having the highest mean. This indicates that the student ratings of teaching effectiveness are positively skewed and may require finer distinctions at the upper end of the scale. This would especially be true if one plans to use a single-item measure of overall teaching effectiveness for summative decisions. For the measure to provide useful information for such decisions, it must be able to discriminate among levels of teaching effectiveness. To insure such discrimination, finer rating scale distinctions may be required, especially at the high end of the rating scale.

One final limitation of the current study is in the single-item validation strategy used. This strategy included convergent validity evidence attained through correlating the single-item measure with the scale and also the interpretation of factor loadings from a factor analysis of the ten (nine scale and one global) teaching
effectiveness items. To further assess the construct validity of the single-item measure, a nomological network could be developed (Schmitt & Klimoski, 1991). This would involve specifying and testing hypotheses about how and to what extent the single-item measure of overall teaching effectiveness relates to other variables. One step in this approach to construct validation would be to relate the single-item measure to a criterion variable. Some of the criteria against which student evaluations of teaching effectiveness have been evaluated in past research are: (1) ratings made by faculty colleagues, (2) ratings made by administrators, (3) faculty self-ratings, (4) ratings made by alumni, and (5) student achievement (Cohen, 1981). Examining the strength of the relationship between these criteria and the student ratings of the single-item measure of overall teaching effectiveness would shed more light on the validity of the single-item measure.

Although one student achievement index (Average Grade of Class) was measured in the present study, it was decided that this criterion was affected by too many factors outside the control on the instructor to provide sound validity evidence. Some of these factors include instructor grading standards, class size, and class variations in student characteristics (e.g. ability). To effectively conduct a validation study of this type, as many factors outside of
the control of the instructor as possible should be controlled (Abrami et al., 1990). No attempts at such control were made in the current study.

Contributions

The present study provides contributions in several areas. First, this study adds to the sparse amount of research on the use of single-item measures. Although these measures are commonly rejected on the basis of poor psychometric properties, little research has been done to either support or refute this contention. This study also presents two techniques that can be used to attain estimates of single-item reliability. This is important because the most common basis for criticizing the use of single-item measures is that they are assumed to have unacceptably low levels of reliability. The results presented in this study and the Wanous et al. (1997) study provide evidence that single-item measures may actually have levels of reliability that are quite acceptable.

The present study also adds to the body of research on student evaluations of teaching effectiveness. In this area, there has long been a debate over the use of multidimensional scales versus global single-item measures to assess instructor effectiveness for the purpose of making personnel decisions. Prior to this study, this debate has
focused almost entirely on the validity of the two measures. This study adds to this literature by further demonstrating the validity of a single-item measure of instructor effectiveness. Another key issue that is given very little attention in this literature is that of reliability. By providing initial estimates of the reliability of single-item measures of teaching effectiveness, the current research provides additional support for the use of single-item teaching effectiveness measures. This study also presents future researchers in this area with techniques that allow the estimation of single-item reliability.

One final contribution of the present study is that the meta-analytic technique is used in a somewhat novel way. Typically, meta-analysis is used to summarize results across basic research studies within a particular domain. In the present study, the effect sizes used for the meta-analysis do not come from different studies, but from subsamples that comprise a larger sample. Some of the key pieces of information available through the use of meta-analysis are the following: (1) an observed weighted mean effect size, (2) an effect size corrected for artifacts, and (3) information about potential moderating variables.
Directions for Future Research

In the research domain of student evaluations of teaching effectiveness, future efforts should focus on the development of a theory of effective teaching. Researchers should look outside of their own research areas to develop such theories. As mentioned in Abrami (1985), information processing theories may provide a framework from which to build a model of teaching effectiveness. The development of these models will allow more effective evaluations of instruction effectiveness, including better student rating forms.

One area of the I/O psychology literature from which teaching effectiveness researchers can draw on is the training literature. Theoretical development in the training literature has led to the recognition that adult learning styles (Newstrom & Lengnick-Hall, 1991) and self-efficacy (Mager, 1992) are key factors that influence training effectiveness. The training literature has also recognized that the evaluation of training programs should occur at multiple levels (Kirpatrick, 1983). In addition to having trainees rate the course and/or instructor, it has become common to assess trainee learning, transfer of learning to the workplace, and business impact of the training program. This four level model of evaluation can
serve as a useful starting point for the development of a more complex model of teaching effectiveness evaluation.

Until such a model of teaching effectiveness is developed, researchers in this area should replicate the present study in different contexts. This would include research at different universities using different student evaluation forms. Because there are many researchers who feel that teaching effectiveness cannot be assessed without considering multiple facets, future research should include the comparison of global single-item measures to multidimensional scales. As described earlier though, results of such studies must be interpreted carefully.

Future research should examine the effect of halo on student ratings of teaching effectiveness. This can be done by including items on evaluation forms that are not conceptually linked to effective instruction. An example would be to have students provide ratings of the physical attributes (e.g. attractiveness, athleticism, etc.) of the instructor. By assessing the relationship between such a variable and teaching effectiveness, one can better establish the impact of halo effect on student ratings.

Based on the positive findings here and in Wanous et al. (1997), future research should be directed at investigating the validity and reliability of other global single-item measures. These investigations should include a
critical assessment of the two techniques of single-item reliability estimation presented here. By further addressing this issue, it is hoped that a technique will emerge as a standard single-item reliability estimation procedure. Once a technique becomes standard and well-accepted, researchers can argue for or against the use of a single-item measure based upon the measure's observed reliability instead of basing arguments on an assumption of low reliability.
# Student Evaluation of Instruction for

**INSTRUCTOR NAME**

<table>
<thead>
<tr>
<th>DEPT</th>
<th>COURSE LEVEL</th>
<th>QUARTER/YEAR</th>
<th>CAMPUS</th>
<th>CALL NO.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**CLASS STANDING**

<table>
<thead>
<tr>
<th>Class Standing</th>
<th>Cumulative Grade Point Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank 1 (Freshman)</td>
<td>3.70 - PLUS</td>
</tr>
<tr>
<td>Rank 2 (Sophomore)</td>
<td>3.00 - 3.29</td>
</tr>
<tr>
<td>Rank 3 (Junior)</td>
<td>2.70 - 2.99</td>
</tr>
<tr>
<td>Rank 4 (Senior)</td>
<td>2.30 - 2.59</td>
</tr>
<tr>
<td>Graduate</td>
<td>2.00 - 2.29</td>
</tr>
<tr>
<td>Graduate Professional</td>
<td>Below 2.00</td>
</tr>
<tr>
<td>Other</td>
<td></td>
</tr>
</tbody>
</table>

**I ENROLLED IN THIS CLASS BECAUSE ...**

- It is specifically required in my major/minor.
- It was one of several choices to meet a requirement in my major.
- It fulfills a GEC/REP requirement.
- It was a free elective choice.

**Instructions**

- Please evaluate the instructor named above and the part of the course taught by that instructor.
- Fill in the appropriate circles completely with a No. 2 pencil.

Evaluate items 1-9 using a scale where the range is from Agree Strongly to Disagree Strongly.

Not Applicable | Agree Strongly | Neutral | Disagree Strongly

1. The subject matter of this course was well organized.
2. This course was intellectually stimulating.
3. The instructor was genuinely interested in teaching.
4. The instructor encouraged students to think for themselves.
5. The instructor was well prepared.
6. The instructor was genuinely interested in helping students.
7. I learned a great deal from this instructor.
8. The instructor created an atmosphere conducive to learning.
9. The instructor communicated the subject matter clearly.
10. Overall, I would rate this instructor as ...
LIST OF REFERENCES


