The Dimensionality and Validity of Student Ratings of Instruction: Two Meta-Analyses

Sylvia d'Apollonia

A thesis in the

Special Individualized Programme

Presented in partial fulfilment of the requirements for the degree of

Doctor of Philosophy at Concordia University, Montreal, Quebec, Canada.

May 1997

0-612-25938-2

ABSTRACT

The Dimensionality and Validity of Student Ratings of Instruction:

Two Meta-Analyses

Sylvia d'Apollonia, Ph.D.

Concordia University, 1997.

Many colleges and universities have adopted student ratings of instruction as one (often the most influential) measure of instructional effectiveness. Although some researchers claim student rating forms are multidimensional, reliable, valid, and uncontaminated by *biasing* variables, other researchers and many instructors continue to express concerns that the validity of summative evaluations based on student ratings are threatened by inappropriate data collection, analysis, and interpretation.

The most commonly used validation design for student ratings is the multisection validity design. Because this validation design has high internal validity and has been used extensively, with many student rating forms under diverse conditions, it provides the most generalizable evidence for the validity of student ratings. However, researchers using this paradigm have reported widely divergent validity coefficients.

Meta-analysis is a useful method of both integrating the findings of a large number of studies and investigating the potential moderating effect of study features. Thus, I conducted two meta-analyses of the multisection validity literature. In the first meta-analysis, I addressed the question. *What is the structure of instructional effectiveness (as judged by students) across student rating forms?*" I concluded that the

forms (at least those used in the multisection validity studies) measure general instructional skill. General instructional skill is a composite of three correlated factors, delivering instruction, facilitating interactions, and evaluating learning.

In the second meta-analysis, I addressed three questions. The first question was, *Are there significant and practically important interactions between moderator variables and the factor structure of student ratings?* The second question was, *What is the overall validity of student ratings as measures of instructional effectiveness?* The third question was, *To what extent is the multisection validity literature consistent, and if it is not consistent, to what extent do study features explain the variability in reported validity coefficients?* The results indicate that there are few interactions between study features and the factor structure of student ratings. They also indicate that there is a medium correlation (.33) between student ratings and student learning. However, methodogical and publication features, quality of evaluation features, student rating form features, achievement measure features, and explanatory features (student, instructor, course and institutional) moderate the validity of student ratings. The presence of these moderators suggests that the student ratings should not be "overinterpreted"; that is, that only crude judgements of instructional effectiveness (exceptional, adequate and unacceptable) should be made on the basis of student ratings.

ACKNOWLEDGEMENTS

I would like to express my deep appreciation to Dr. Philip Abrami, my thesis advisor, who has challenged and guided me throughout this research. I am also indebted to the members of my research committee, Drs.Gary Boyd and Ron Smith for their support and encouragement. I would also like to thank Drs. Peter Cohen, Ken Feldman, Herbert Marsh, and the other members of the American Educational Research Association, Special Interest Group Faculty Evaluation. I have benefited not only from their ground-breaking work but also from their constructive criticism of my work.

The research would not have been completed without the help of the research assistants who helped code the studies. I therefore would like to thank Magda Hechema, Gretchen Lowerison, Bruce Netherwood, and Sean Spurvey for their diligence. Similarly, I would like to thank my colleagues, Helena Dedic and Steven Rosenfield, who spent long hours with me discussing the implications of this research. I also would like to thank Amanda Foster, who helped assemble the final document.

Finally, I would like to thank my family and friends who patiently gave me emotional support throughout this process.

TABLE of CONTENTS

LIST OF TABLES AND FIGURES

INTRODUCTION

The historian of science. Derik de Solla Price (1975). described the following three stages in the growth of a scientific discipline:

- an early slow growth phase during which the seminal precursor papers are published;
- an exponential growth phase during which there is an exponential increase in published work; and
- a declining growth phase during which the field is saturated.

In North America. student evaluation of instructors was introduced at major post-secondary institutions (Harvard. Purdue University. the University of Texas. etc.) in the mid - 1920's (Marsh. 1987). The approximate number of published and unpublished studies produced per year during the first five decades of the century was 3 ; during the decade from the mid- sixties to the mid -seventies. 28; during the decade from the mid seventies to the mid-eighties. 132; and. during the decade between the mid-eighties and mid-nineties. 72 [1]. Although. this informal analysis underestimates the number of studies on student ratings. it illustrates that research in the field of student evaluation is entering the third stage. Although the primary research is declining: researchers are now attempting to integrate this large complex body of research literature [2] and to make sense of the phenomenon.

[1] Data to 1985 was obtained from Marsh (1987) using the key words *students' evaluations of teaching*. Subsequent data was obtained from ERIC using the same search terms.

[2] Jackson (1980) defines an integrative review as one in which the reviewer has inferred generalizations (*e.g.* laws) about a substantive phenomenon from a set of studies directly bearing on the phenomenon.

Despite much lip-service to the importance of this integration. much of the social sciences can be characterized as having a "relatively unimpressive degree of cumulative knowledge " (Feldman, 1971, p. 86; see also Glass. 1976: Light & Smith. 1971; Meehl, 1978); thus, hampering both theory construction and policy decisions. Koch's (1981) examination of the past century of psychology lead him to conclude that such cumulativeness did not occur; fractionatiön rather than integration was the rule. Critics often consider that this non-cumulativeness is a characteristic only of the *soft sciences*. However. Hedges (1987) compared the cumulativeness of data from the physical sciences to that from the behavioral sciences and concluded that behavioral science research was not substantively less cumulative than physical science research. Whenever a large number of observations is taken. diversity arises. It is the task of the integrative reviewer to summarize the diverse findings into a coherent theory and to explain the diversity in terms of experimental characteristics.

As mentioned above. there is a very large body of research on students' evaluations of instruction carried out since the mid- 1920's. Student ratings of instruction are widely used in post-secondary institutions to assess the effectiveness of instruction. They are used to aid students in course selection. to provide instructors with feedback for course and instructional improvement. to provide researchers with information on the teaching-learning process. and to provide administrators with information for hiring and tenure decisions.

Because performance ratings, of which student ratings are one type. are used to make personnel decisions, concerns have frequently been expressed that such ratings are

" subjective, ... biased, and at worst purposefully distorted" (Saal, Downey, & Lahey, 1980, p 413). Thus, organizational and personnel psychologists have frequently investigated the quality of performance data (*e.g.*, American Educational Research Association, American Psychology Association, & National Council on Measurement in Education, 1985; Cooper, 1981; Cronbach, & Meehl, 1955; Saal, Downey, & Lahey, 1980; Thorndike, 1920; Thurstone, 1937). Historically, there have been two parallel approaches to the study of the quality or veracity of ratings. One approach adopts the construct validity approach (Cronbach, & Meehl, 1955) whereby high inter-rater reliability, high convergent and divergent validities, and low method effects are indicators of high quality. These methods usually employ the multitrait-multimethod procedures popularized by Campbell and Fiske (1959). Another approach adopts the position that the absence of such rater errors and biases as range restriction (leniency, central tendency, and severity), halo, inter-rater disagreement, and unstable dimensions are indicators of the high quality of ratings (Murphy, & Balzer, 1989; Saal, Downey, & Lahey, 1980).

In the area of student ratings of instruction, both approaches have been used to investigate the reliability, dimensionality, and validity of these performance ratings (*e.g.*, Abrami, Cohen, & d'Apollonia, 1988; Abrami, d'Apollonia, & Rosenfield, 1996; Cohen, 1981; Feldman, 1977, 1978, 1989; Marsh, 1984, 1987; Murray, 1980). Although there is general agreement that student ratings are reliable, controversy persists concerning their dimensionality and validity. Therefore, the goal of this thesis is to address these issues by conducting two meta-analyses; one on the factor studies of student ratings of instruction, and another on their validity.

This thesis is therefore divided into three major parts. The first part (PART I) consists of a literature review of meta-analysis. emphasizing the steps that are required to conduct a meta-analysis that is free of threats to the interpretation of its findings.

The second part (PART II) includes a literature review of the dimensionality or structure of student ratings, emphasizing the studies that have attempted to determine the dimensionality across forms, the controversies surrounding the interrelationships among scales, especially between global and specific scales. and the disagreements among researchers concerning which scales are appropriate measures for personnel decisions. PART II also includes a description of the methods I used to integrate the factor studies of student ratings, the results of this meta-analysis (*i.e.*. a common factor structure *across* forms), and a discussion of the implications of the structure. so revealed. to the use of student ratings as measures of instructional effectiveness.

The third part (PART III) includes a review of the literature on the validity of student ratings. critically examining definitions of validity. four common validation designs (multitrait-multimethod studies. laboratory studies. studies investigating the absence of *biasing* factors. and multisection validity studies). and prior integrative reviews of multisection validity studies. It also includes a description of the methods I used to conduct the meta-analysis of the multisection validity studies. the results of the analysis. and a discussion of these results. emphasizing the influence of study characteristics on validity.

Finally. I will summarize the conclusions to be drawn from the two meta-analyses, and present the bibliography and appendices.

PART I

META-ANALYSIS: THE INTEGRATION OF RESEARCH FINDINGS

Literature Review

There have been a number of criticisms of the traditional research review (Feldman. 1971; Glass. 1978; Jackson. 1980: Pillemer. & Light. 1980; Slavin. 1984) with respect to representativeness. reliability and replicability. These shortcomings arise because selection criteria. weighting of evidence. and interpretive biases are often implicit rather than explicit. Some of the above problems can be eliminated by the application of scientific rigour to the review process (Cooper. 1982; Feldman. 1976: Jackson. 1980; Light, & Pillemer, 1982; Slavin. 1986). However; traditional narrative reviewers rarely report their mathematical and inferential procedures nor use measures of treatment magnitude (Cooper, 1981). Rather, traditional reviewers often assess the overall conclusions and consistency of a set of studies by tallying the number of studies that report significant findings versus the number of studies that report non-significant findings (Light & Smith. 1971; Meehl. 1978). This vote counting method has been extensively criticized (Glass. McGaw. & Smith. 1981; Hedges. & Olkin. 1980: Hunter. & Schmidt. 1990: Hunter. Schmidt. & Jackson. 1982) as not taking into consideration the limitation of significance tests. These tests were developed to control for Type 1 errors (*i.e.*. to limit the probability of inferring an effect in the absence of a *true* effect to 5%). However. they do not control for Type II errors. That is. the probability of inferring the absence of an effect in the presence of a true effect can be as high as 95%. Traditional reviewers treat both types of errors as if they were equal when the use vote counting to summarize a set of studies.

In the 1930's, statistical methods originating from agricultural and astronomical studies were developed to combine the results from a series of independent studies (Cooper, & Hedges, 1994; Hedges, & Olkin, 1985; Kulik, & Kulik, 1988). However, these methodologies were not in general use in the social sciences until researchers elaborated these methods into a set of integrative research reviews, labelled *meta-analysis* by Glass (1976). These approaches to the integrative research review are of two[3] types:

- parametric tests in which the treatment effects from individual studies that form a homogeneous set are transformed to a common metric and combined; and,
- non-parametric or omnibus tests in which the probability values of individual studies are combined.

Since in this thesis I was interested in the degree to which student ratings of instruction predicted student learning across studies, and not the cumulative significance of the research, I combined treatment effects across studies, not probability values. I therefore will only review research based on the former approach.

Historical Development of Meta-analysis

The aggregation of experimental results from different studies in the physical and biological sciences was more common since, unlike the situation in the social sciences, measurements are usually uniscalar (Hedges, & Olkin, 1985). Therefore, many of the

[3] A third category, partially-parametric tests in which vote-counting estimators are used to estimate effect sizes or effect magnitudes, is described by Hedges and Olkin (1985).

early statistical methods of aggregating probability values (Fisher, 1932; Pearson, 1933; Tippet, 1931) or means (Cochran, 1937; Yates, & Cochran, 1938) were derived for agricultural research. These methods, especially combining probability levels, were introduced into the social sciences by Jones and Fiske (1953), Mosteller and Bush (1954) and Wilkonson (1951).

Glass (1978) popularized the use of quantitative review techniques in the social sciences by introducing the use of *scale-free* measures (*eg*., effect size, effect magnitude). At the same time, Hedges (1981), Hunter, Schmidt, and Jackson (1982), and Rosenthal (1978) developed other approaches to quantitative review synthesis (sampling-variance weighted, validity generalization, and combined probability approaches, respectively). In addition, Hedges (1981) and Hunter, Schmidt, and Jackson (1982) elaborated on Glass' approach by developing correction procedures for bias, unreliability and range restriction. Recently, statisticians have begun to develop multivariate approaches to meta-analysis (Gleser, & Olkin, 1994; Hedges, & Olkin, 1985, Raudenbush, Becker, & Kalaian, 1988). In addition, there are now a number of books that describe and compare the different meta-analytical techniques (Cooper, & Hedges, 1994: Hunter, & Schmidt, 1990; Wolf, 1986).

Light and Pillemer (1982) suggested that meta-analysis was especially needed in the area of policy decisions. Recently, in the United States, legislation was passed requiring that guidelines on health care policy research be based on systematic research integration, such as meta-analysis (Cooper, & Hedges, 1994).

Not only has meta-analysis become a well-established analytical tool, it has now become a research topic of its own. Feldman (1971) called for the consideration of the integrative review process as a research topic, *per se*. Jackson (1980) analyzed different research procedures using the criteria of primary research and recommended that the rigour of primary research be applied to the integrative review process. Cooper (1982) applied models of internal and external validity from primary research (Campbell, & Stanley, 1966; Cook, & Campbell, 1979) to meta-analysis. Finally, meta-analysis has been used to construct and test causal models (Cook, Cooper, Cordray, Hartmann, Hedges, Light, Louis, & Mosteller, 1992).

Description and Criticisms of Meta-analysis

Glass described meta-analysis as an approach to conducting an integrative review that makes use of statistical tools both to aggregate the summary statistics of a large set of studies on a given topic and to explain the variability in research findings (Glass, McGaw, & Smith, 1981). There are five steps in conducting a typical meta-analysis: problem formulation (specifying inclusion criteria), locating studies, calculating individual outcomes, coding study features, and data analysis including model testing and interpretation (Abrami, Cohen, & d'Apollonia, 1988).

Kulik and Kulik (1988, 1989) described four features which, in their view, characterize meta-analysis. They are:

- Meta-analysis is the analysis of the research literature; it is not itself a primary study.
- Meta-analysis covers a *large* body of literature, not a handful of studies.
- In a meta-analysis, summary statistics are aggregated, not probability values. Thus, a

meta-analysis provides information on the direction and magnitude of the effect. not only on the significance.

- The focus of a meta-analysis is to attempt to explain the variation in findings.

Not all meta-analyses include all four characteristics. For example. only the first two features were present in the published meta-analysis on the validity of student ratings by Abrami (1984). and only the first three in those by Dowel and Neal (1982). Feldman (1989) and McCallum (1984). Cohen (1980. 1981. 1982. 1983) was the only reviewer of the multisection validity studies who included all four features. These five steps are briefly described below (Cooper. & Hedges. 1994).

Problem Formulation

This stage of the procedure refers to the specification of the phenomenon in question by defining the criteria for study inclusion. Thus. it includes operationally defining the constructs and establishing criteria for the conceptual domain of inquiry. the permissable statistics. and methodological quality. Most of the critics of meta-analysis fault it for having too liberal inclusion criteria. There are two main concerns. the *apples and oranges* problem (Glass. 1977. p 356) and the *garbage in garbage out* problem (Eysenck. 1978. p 517). That is. there are concerns about both the relevance or commensurability of included studies and the influence of the inclusion of poor quality studies the conclusions reached in the meta-analysis.

Smith and Glass (1977), in the study that launched meta-analysis, defined the conceptual domain, the effectiveness of psychotherapy, very broadly. Their study was criticized for producing conceptual confusion by ignoring distinctions between different therapies (Eysenck. 1984; Presby. 1978). Cooper (1979) argued that the aggregation of studies is only merited when they share common hypotheses or common operational definitions. Similarly, Slavin (1983) objected to the scope of the meta-analysis on goal-structures carried out by Johnson. Johnson. and Maruyama (1983). Slavin (1984. 1986, 1987) also argued that the criteria for inclusion should be the germaneness rather than only the methodological quality of a study. The above critics argue that it is logically inappropriate to aggregate across studies which use disparate operational definitions. subjects. settings. and measures (Wolf. 1986); that is. they argue that meta-analysis is analogous to mixing *apples and oranges.*

Proponents of meta-analysis. on the other hand. argue that aggregating across studies is not logically different from aggregating across subjects within a study. Moreover. they contend that to restrict integration to studies that are the *same* is trivial as only *different* studies require integration (Glass. McGaw. & Smith. 1981). They also contend that the issue of the *possible* incommensurability of studies is better investigated empirically by coding the studies and testing for heterogeneity rather than decided *a priori.*

The purpose of the review should dictate the breadth of the problem definition. Obviously the breadth of the problem definition will reflect the reviewer's purpose. There have been a number of criticisms of meta-analysis for focusing on main effects at the expense of interactive effects (Cook. & Leviton. 1980; Light. 1987; Slavin. 1983). These critics maintain that an integrative reviewer must not only describe the variability in findings, but also explain it. Thus. meta-analysts should also investigate possible interactions due to differences among subjects. settings. locales and contexts. Perhaps this is not so much a criticism of the reviewer's method as it is a criticism of the reviewer's purpose.

Perhaps the most contentious issue. in the problem formulation stage. is the inclusion of studies of *questionable* quality. Glass argued that it is inappropriate to discard studies on the *a priori* basis of poor design because studies with many flaws may give the same results as *perfect* studies. Critics (Eysenck. 1984; Slavin. 1984) argue that aggregating studies of questionable validity results in questionable meta-analyses. The effects of this indiscriminate inclusion. they argue. can affect not only the estimation of the mean effect size but also the estimation of the variance among studies. Furthermore. such threats to internal validity (Campbell. & Stanley. 1966). do not necessarily cancel each other out; in certain domains. these threats are likely to bias the results in one direction (Slavin, 1984). This issue was coined the "*garbage in - garbage out*" problem by Eysenck (1984). Most researchers (Kulik. & Kulik. 1988; Strube. & Hartman. 1982; Wolf. 1986) consider this to be an empirical question and suggest that meta-analysts include all studies. assess their methodological quality. and either weigh the studies by methodological quality (including by 0. *i.e.*. excluding them) or measure the impact of methodological quality on their conclusions.

Slavin (1984) agreed that meta-analysts could . *in principle*. control for the inclusion of studies of questionable quality; however, he stated that, *in practice*, this is not done. Although meta-analysts report that methodological issues significantly interact with outcomes, they still interpret overall effect sizes or magnitude. Slavin points out that utilizing the results of a homogeneity test to limit interpretation to homogeneous data sets would resolve this problem. However, most meta-analysts persist in interpreting the mean effect size of heterogeneous data-sets. For these reasons, Slavin (1986,1987) developed what he calls *best-evidence* synthesis in which he called for the exclusion of studies that are not explicitly germane to the problem statement, contain threats to internal and external validity, and used very small sample sizes. He also stated that pooled effect sizes should only be computed (and therefore interpreted) for homogeneous data sets.

To summarize, a meta-analyst should state the research question as specifically as possible and define relevant variables. He or she should clearly describe and justify both inclusion and exclusion criteria. I do *not* recommend excluding studies on the basis of methodological quality and therefore, do not recommend Slavin's *best-evidence* synthesis. Although meta-analysts should exclude erroneous data (typographic errors, *etc.*), they should include all studies and code for methodological weaknesses along multiple dimensions (*e.g.*, random assignment, reactivity, *etc.*). They can subsequently assess the impact of such methodological flaws on the mean effect size and on the distribution of effect sizes. If necessary, they can subsequently justify the elimination of outliers. In addition, they should use the homogeneity test (Hedges, & Olkin, 1985) to test the homogeneity of a date set or subset that has been pooled. Data sets that are heterogeneous should be stated to be heterogeneous and the pooled effect size interpreted with extreme caution.

Locating Studies

The second step in a meta-analysis is to exhaustivly search the literature for empirical studies, both published and non-published. The findings reported in the studies are the subjects in a meta-analysis. The reviewer, like the primary researcher, attempts to analyze a sample from this population of elements. Sampling, in meta-analysis is most similar to survey-sampling (Glass, McGaw, & Smith, 1981, p 24), and is "apt to be non-random and biased" (Strube, & Hartman, 1982, p 133). Thus, the best defence against systematic bias is to locate as large a set of relevant studies as possible and to document the search strategy. Since detailed procedures for carrying out such literature searches have been published (Glass, McGaw, & Smith, 1981; Hunter, & Schmidt, 1990; Hunter, Schmidt, & Jackson, 1982; Rosenthal, 1994), I will only discuss this step briefly.

Cooper (1984) categorized search strategies according to three channels, informal, primary and secondary. Informal channels include the reviewer's own research, the *invisible college* that connects certain researchers but excludes others, and conventions. Primary channels include the studies present in the reviewer's personal collection as well as the bibliographies present therein. Secondary channels consist of the publicly available research and includes abstract and indexing services, computer searches, bibliographies and public library collections. Extensive reliance on the first two channels produces systematic biases in the direction of the reviewer's expectations (Cooper, 1984), and is likely to reduce the reliability of the literature search. For this reason, some reviewers have recommended that reviewers limit themselves to the secondary channel. However most meta-analysts recommend using *all* available sources in order to get the most comprehensive set of studies possible.

Most literature searches today rely heavily on computer searches. However, overreliance on this technique at the expense of manual searches can also pose some problems. The computer data bases are biased to North American studies usually published in English (Kelly, 1986). The effectiveness of computer searches depends on the exact specification of a search strategy via key-words. Unless this is reported, the meta-analysis is not replicable. Furthermore, different data-bases access different journals. Since one of the goals of meta-analysis is to make the review process replicable, it is essential that the exact search strategy that is used be reported.

There are a number of practical concerns that arise after one has completed a literature search: too many studies, missing studies, and a technologically biased set of studies. In certain domains, a complete search of the literature will result in the retrieval of thousands of studies, many of which are unpublished theses. How does one begin the Herculean task of acquiring, reading, and coding every single study? Light and Pillemer (1982) suggests three additional options to the inclusion of every study. These are to stratify the studies and select a random sample from each strata, to use only published studies, or to use a panel of experts to generate a list of studies for inclusion. The latter two options, however are likely to introduce serious biases and are therefore not recommended.

Missing studies can produce systematic bias because of the greater value of statistically significant findings than of non-significant findings to both researchers and editorial boards. This has been called the *file drawer* problem by Rosenthal (1978).

Given that this problem exists (Greenwald, 1975; Kraemer, & Andrews, 1982), reviewers have made two recommendations:

- Calculate and report the Fail Safe N, the number of additional studies that would be necessary to reverse a conclusion that a significant relationship exists, (Orwin, 1983; Rosenthal, 1979).
- Estimate the effect of censoring rules, *i.e.*, excluding non-significant findings, in order to evaluate the plausibility of publication bias (Hedges, & Olkin, 1985).

I would therefore recommend that meta-analysts completely describe their search strategy. Some reviewers, in an attempt to insure that other researchers would be able to replicate the search strategy and study retrieval, limit the search strategy to electronic data bases and branching from studies retrieved in this manner. However, this insures replicability at the expense of representiveness; that is, reliability at the expense of potential bias. Therefore, I would recommend that the meta-analyst use all available sources, both formal, and informal, to identify potential studies. In addition, the meta-analyst should also determine the probability that some studies have been systematically missed. If this *file drawer* problem is likely, the Fail Safe N should be calculated.

Calculating Individual Outcomes

The third step is the extraction of the summary statistics from the studies and, if necessary, their transformation to a common metric. There are a number of issues that need to be addressed during this step. Firstly, the meta-analyst must decide on the number of outcomes to extract from each study thus dealing with the issue of non-independence. Secondly. the meta-analyst must attend to extracting outcomes reliably. Thirdly. the meta-analyst must decide how to compute summary statistics.

Non-independence problem. The first decision that needs to be made is the number of outcomes to extract from each study. Is it to be every outcome reported in a study. one outcome from each study. or some intermediate number? This issue raises the problem of non-independence since the presence of multiple data points from a single study introduces "complicated patterns of statistical dependence" (Glass. McGaw. & Smith. 1981, p 200. see also Abrami. Cohen. & d'Apollonia. 1988; Hedges. 1986: Landman. & Dawes. 1982). Statistical dependencies can arise from a number of sources:

- multiple measures across the same subjects both within the same study or among studies (repeated measures and multivariate designs):
- multiple investigations using different subjects within one study, (including factorial designs); and.
- multiple studies carried out by the same research team.

Each situation introduces dependency among outcomes because of similarities among settings. measures. subjects. *etc.*

Non-independence is a problem because most significance tests include an assumption that the data points are independent. Assumption violation can lead to serious inflation of Type I and Type II error rates (Glass. McGaw. & Smith. 1981: Raudenbush. Becker. & Kalaian, 1988; Reeves 1989; Rosenthal, 1995). For example, Monte Carlo studies have indicated that non independence inflates the Type I error rate from the nominal .05 to .15 (Reeves, 1989). That is the Q_T statistic is too liberal when data are not independent. Glass and his colleagues also demonstrated that non-independence can increase the standard error about the mean effect size by a factor proportional to the number of studies. For a meta-analysis containing fifty studies, the true standard error is 30 times greater than that calculated assuming independence. For this reason, Glass (Glass, McGaw and Smith, 1981) did not recommend inferential statistics in meta-analysis since the increase in error may cancel out the advantage of increased power. However, few scientists are prepared to abandon inferential statistics.

Since sampling within meta-analysis is analogous to survey cluster sampling (McGaw, 1988), some meta-analysts (Glass, McGaw, & Smith, 1981; McGaw, 1988), have suggested using cluster analysis (Kalton, 1983; Kish, 1965; Sudman, 1976) to estimate the effects of ignoring non-independence. In this way, if the effects of non-independence are trivial, they can be ignored; on the other hand, if the effects are significant, the error rates can be adjusted appropriately.

Other researchers have reported that non-independence may not be a problem *in practice* (Center, Skiba, & Casey, 1986; Landman, & Dawes, 1982). This position is supported by Monte Carlo studies (Tracz, & Elmore, 1985) that show that non-independence does not affect the Type I error rate of the effect magnitude (both *r* and *z*). Rosenthal and Rubin (1986) also found that non - independence produces conservative estimates of the mean and median effect sizes. Hattie and Hansford (1984) compared their conclusions when they ignored non-independence to their conclusions when they corrected for it by jackknifing and found trivial differences. They therefore ignored non-independence `in estimating parameters. but eliminated non-independence by jackknifing when using inferential statistics.

Nevertheless. because of the potential inflation of error rates, some analysts recommend selecting one outcome per study. thus choosing the study as the unit of analysis (Kulik. 1983; Kraemer. 1983: Mansfield. & Busse. 1977) and avoiding the problem of non-independence (Bankgert-Drowns. 1986). There is. however. no agreement on how this selection is to take place. Is it by averaging across all similar outcomes[4]. by averaging across reliable outcomes. defined by interrater agreement. outcome reliability. or consistency rules. (Matt. 1989). by random selection. or by weighting the multiple effect sizes and calculating a composite mean effect size. Different methods of weighting have been proposed. Some researchers have weighted by the inverse of the number of effect sizes per study (Johnson. Johnson. & Maruyama. 1983; Kendall, 1979) or by Tukey's jackknife procedure (Glass. McGaw. & Smith, 1981; Hattie. & Hansford. 1984). However. these methods have the drawback that they do not take into consideration the interdependencies among multivariate outcomes.

When the study reports findings from multiple dependent measures that are conceptually distinct, some meta-analysts (Cohen. 1981. Rosenthal. & Rubin. 1978) recommend selecting one finding per dependent measure and conducting separate meta-analyses for each dependent measure. This avoids the problem of non-independence but makes it difficult to analyze effects across measures. For this reason. most meta-analysts

[4] That is all outcomes representing the same construct.

who are interested in modelling a phenomenon, rather than describing a set of studies, do not recommend this procedure (Abrami, Cohen, & d'Apollonia, 1988, Becker, 1992; Raudenbush, Becker, & Kalaian, 1988).

However, reviewers may wish to integrate findings across multiple dependent measures within a study. In such cases, Rosenthal and Rubin (1986) have suggested extracting one finding per each dependent measure and weighting each multiple effect sizes within each set by the intercorrelations among outcomes and the degrees of freedom per study. This takes into consideration both the sample size for each study and multiple outcomes due to multiple dependent measures. However it does not deal with the non-independence due to multiple outcomes for the same dependent measure. Thus, it does not allow for within-study comparisons (Abrami, Cohen, & d'Apollonia, 1988).

Finally, other researchers, i.e., Abrami, Cohen, and d'Apollonia (1988); Glass, McGaw, and Smith (1981), and Raudenbush, Becker and Kalaian (1988) pointed out that collapsing multiple outcomes, whether between groups or within groups, into one average outcome, obscures important questions about differences across measures and study features. Moreover this solution neither removes all non-independence, nor more importantly, does it allow for generalizations across meta-analyses (Raudenbush, Becker, & Kalaian, 1988). These researchers recommend choosing all findings and modelling the interdependencies by using multivariate rather than univariate approaches (Gleser, & Olkin, 1994; Hedges, & Olkin, 1985; Raudenbush, Becker, & Kalaian, 1988). This analytical solution will be discussed on page 33.

In conclusion, there are three proposed solutions to the problem of non-independence (when there are both multiple dependent measures and multiple findings per dependent measure):

- Ignore the problem of non-independence due to both multiple findings and multiple dependent measures and select all findings from the studies. Estimate the influence of non-independence on sampling variance and if significant, adjust the error rates accordingly.
- Avoid the problem of non-independence due to multiple dependent measures by selecting only one finding per dependent measure per study, averaging findings within dependent measures, or calculating a weighted composite effect size for the combination of multiple dependent measures. Conduct separate analyses for each dependent measure.
- Deal with the problem of non-independence due to multiple dependent measures by selecting all findings within studies but analyze the data multivariately.

Reliability problem. Whether the decision is to extract all outcomes or only a subset, meta-analysts must stipulate how to extract outcomes. Abrami, Cohen, and d'Apollonia (1988) in a review of the seven meta-analyses conducted on multisection validity studies reported large discrepancies in the number of outcomes extracted from common studies. For example, there is only 55% agreement between Cohen (1981, 1983) and McCallum(1984) for *Overall Instructor* rating and only 25% for *Overall Course* rating.

Matt (1989) also reported a very low interrater reliability in the number of effect sizes extracted from the same subset of the psychotherapy literature between Smith, Glass and Miller (1980) and the three coders in his study using a conceptual redundancy rule. This rule, originally proposed by Smith and Glass, excludes all effect sizes which do not add any incremental validity. Matt proposed three other rules for outcome selection; *coder agreement, outcome reliability* and *outlier truncation*. The *coder agreement* rule only includes nonredundant effect sizes on which coders can agree. The *outcome reliability* rule only includes nonredundant effect sizes that were measured reliably. The *outlier truncation* rule excludes nonredundant effect sizes if they exceed a fixed outlier limit. However, the application of these three additional decision rules did not remove all judgement bias.

The problem of the reliable extraction of outcomes is increased when factorial designs are included in the meta-analysis. For example, which and how many contrasts will be extracted? The number of possible contrasts (k) in a factorial design can be calculated from the number of cells (n) in the design.

$$k = \sum_{i=1}^{n-1} i$$

Thus, in a study with a 2x3 factorial design, there are 15 possible contrasts when contrasting only single cells. With a 2x3x2 design there are 66 possible contrasts. If one also includes contrasting combination of cells (*i.e.*, row and column means) the number of possible contrasts is even greater. Obviously, not all contrasts are salient; it depends on the question(s) being addressed. However, a search of the literature did not reveal any discussion on how this decision (to include the maximum or the minimum number of contrasts) is to be made.

Choice of common metric. The two classes of common metrics used in meta-analysis are effect size (used to compare treatment effects) and effect magnitude (used to compare the magnitude of the relationship between two variables). Effect size is the standardized mean difference (d) between two treatments; while, the effect magnitude is the product moment correlation (r) or Fisher transformation (z). There are procedures for the conversion of most summary statistics to a common metric: *i.e.*, from definitions (*e.g.*, r, z, d), from significance tests (*e.g.*, t, F), from significance levels (*e.g.*, .05, .01, not significant), and from other effect sizes (*e.g.*, r, d, g). These have been described at length (Cohen, 1977; Glass, McGaw, & Smith 1981; Hedges, & Olkin, 1985; Rosenthal, 1984; Rosenthal, 1994), and therefore will not be discussed here. However, concerns remain on the use of *r* as opposed to *z*, on the commensurability of the different metrics; and on estimated *versus* exact calculations of transformations (Cooper, 1981; L'Hommedieu, Menges, & Brinko, 1987). There is also a debate about the influence of sample size on some metrics (Strube, 1988). If design effects (*eg.* sample size) are correlated with some metrics (omega squared) and not others (ES), the interpretation of moderators will vary with the metric of choice. Strube (1988) suggests that when such design effects contribute to the variance in outcomes, design differences "must be considered as a serious alternative explanation to any substantiative moderators" (Strube, 1988, p. 344); especially when the metric of choice is omega squared.

In conclusion, the meta-analyst must decide how to deal with the presence of multiple findings within a study. I recommend that all findings that are relevant to the problem statement be extracted and that the dependencies be explicitly modelled, using the multivariate techniques developed by Hedges and Olkin (1985) and Raudenbush, Becker, and Kalaian (1988). The selection of all findings, rather than a subset, will also increase the reliability of data extraction. Nevertheless, decisions on which findings to exclude, and on computation procedures should be explicitly described and defended since the findings become the dependent variables in subsequent models which attempt to explain variability in findings. Any question on their reliability seriously hampers the testing of these models.

Coding Study Features

Meta-analysis is most useful in those areas in which findings are inconsistent. In such cases, the purpose of the analysis is to explain (or at least to predict) study outcomes from study characteristics. That is, the objective of the meta-analysis is to explain the variability in the phenomenon of interest, not merely to summarize research findings. The fourth stage in a meta-analysis, coding study features, is required to analyze this variability. It includes developing a coding schema, coding the studies, deciding how to handle missing data (study features), and judging the adequacy of coding.

Developing a coding schema. The study characteristics become, in effect, the explanatory variables in a model explaining the variability in findings. Thus, specification errors (both of commission and omission) seriously jeopardize the success of this model (Pedhazur, 1982). For example, when relevant variables correlated to the variables present in the model are omitted, the estimation of regression coefficients is biased. If the omitted variables are not correlated to the variables present in the model, the estimation of the regression coefficients is not biased; however, the standard errors of the regression coefficients are increased reducing the power of statistical tests. Similarly, the inclusion of irrelevant variables does not bias the estimates of the regression coefficients; however it reduces the sensitivity of significance tests . Hunter and Schmidt (1990) also point out that the addition of unnecessary variables increases the chances of spurious significant effects. They suggest that variables only be included if there is theoretical justification for their inclusion. A less conservative view is proposed by Stock (1994) who suggests that the meta-analyst first formally speculate on the size and direction of the influence of each study feature. The decision to include a study feature should be based on the questions being addressed, the prevalence of descriptions of the variable in the literature, coding reliability, and the associated costs of coding the variable. Since these variables (and their interrelationships) are the embryonic hypotheses concerning the factors which affect the size of the treatment effect, this selection is enhanced by the prior construction of a conceptual model of the phenomenon in question.

Although all meta-analysts describe coding study features as one of the stages in a meta-analyses and most describe methods of measuring inter-rater reliability, few have explicitly described procedures to develop such a schema. In most cases, study features appear to be selected on the basis that other researchers have selected them. This problem of variable selection is increased in the educational literature because of a dearth of theoretical models. For example. Borich (1977) state that "in the field of teacher behaviour, however, persuasive theories providing a logical coherent rationale ... have not been forthcoming. This has been perhaps the greatest weakness of the voluminous research...which empirically examines relationships between teacher and pupil behaviours" (Borich, 1977. p. 10).

Borich (1977) suggests methods of developing a valid system of evaluating teacher effectiveness. The first step is to search the literature for significant relationships and rationally select promising behaviours and skills. The second step is to build a nomological network indicating antecedent. intervening and terminal behaviours, to test the validity of the above relationships. and to sequentially order the behaviours and skills. The third step is to construct a taxonomy (a model or hierarchical representation of the relationships among behaviours) emphasizing the important distinctions and minimizing the superfluous ones. This model can be subsequently tested via CFA (Confirmatory Factor Analysis) or LISREL (Linear Structured Relationships) (Hill. 1984; Marsh. 1991a). Thus. three stages are indicated; selecting variables on the basis of the literature. chunking variables on the basis of relationships. and constructing a hierarchical representation or model on the basis of theory.

Most reviewers do not explicitly describe how they select and define the study features. Thus, it is difficult to judge the adequacy and comprehensiveness of their analysis of outcome variability. Abrami *et al.* (1988, 1990) have described a method of selecting and coding variables which is both inclusive and systematic. Essentially, it consists of extracting all the variables in the relevant literature and classifying them on the basis of three dimensions. The first dimension includes a framework describing the study features. This framework was modified and forms the basis of the coding of study features described in this thesis. The second dimension describes how the variables in question were treated in the primary study. For example, did the researcher experimentally investigate the influence of the variable, statistically control for its effects, or simply discuss possible effects. The third dimension describes the source of the coding. For example, in some studies the primary researcher did not comment on certain variables; however, the reviewer can extract and code these variables. This nomological coding, especially the first dimension, has been an effective tool to analyze the adequacy of the selection of potential explanatory variables by prior integrative reviewers of the multi-section validity literature (Abrami, & d'Apollonia, 1990).

Special attention should be given to developing a coding schema which captures methodological quality since the aggregation of studies of variable quality may both bias the mean effect size or magnitude and increase the confidence interval (Slavin, 1983, 1984, 1987). A number of researchers have described both methods of assessing methodological quality and weighting on the basis of methodological quality (Hall, Tickle-Degben, Rosenthal, & Mosteller, 1994; Rosenthal, 1984; Summers, 1989; Wortman, 1994). However, there are a number of problems with this solution:

- Jackson (1980) and Bryant, & Wortman (1984) have pointed out that evaluating the impact of methodological quality requires that a sufficient number of *good* studies be present to provide a comparison. Not only is the decision on which studies are *good* potentially biased; but, there may not be a sufficient number of *good* studies to make an appropriate comparison.

- Weighting by methodological quality assumes that the relationship between the methodological characteristic and the outcome is linear. This is not necessarily so. For example, instrument length influences the reliability of a test as a measure of student learning. However, both very short and very long forms are likely to adversely affect the test characteristics, and therefore the influence of this methodological variable will be curvilinear. A researcher conducting a meta-analysis on the relationship between student ratings and student learning would need to weigh studies on the basis a nonlinear equation representing the influence of test length. Moreover, differences in sample size (number of studies) along the methodological variable (*e.g.*, test length) may preclude extrapolating the influence of study quality (Slavin, 1984).
- Methodological quality is multidimensional and not necessarily causally related to outcomes. L'Hommedieu, Menges, and Brinko (1987), in a meta-analysis on the effectiveness of student rating feedback on college instructors, attempted to code for methodological quality using the traditional indices of research rigour. They found that although they could assess the quality of a study on any one dimension, they could not reliably [5] and accurately translate this into a global summative rating. In

[5] The inter-rater reliability of methodological quality is often quite low (Hattie & Hansford, 1984).

this case, multivariate descriptions of methodological quality may be more instructive than one summative score. They concluded that they would neither weight by a single global quality index, nor trust the conclusions of meta-analyses that did. Shadish and Haddock (1994, p 265) also concluded that "we cannot recommend that other weighting schemes (*i.e.*, for quality) be used routinely prior to combining results".

Wortman (1994) described a strategy for coding study quality that was based on two validity approaches. The first was based on the threats to validity in true- and quasi-experimental designs (Campbell, & Stanley, 1966; Cook, & Campbell, 1979).The second was based on the threats to validity in the randomized controlled trials used extensively in clinical trials in medicine (Sacks, Berrier, Reitman, Ancona-Berk, & Chalmers, 1987). Wortman (1994) has integrated these two approaches and grouped these threats as affecting the following four types of validity: construct, statistical conclusions, internal, and external. Wortman (1994) recommended coding studies on the basis of threats to validity and subsequently conducting a three-stage triage process in which

- studies are excluded (given a weight of 0) if they are not relevant to the problem formulation on the basis of fatal threats to construct and external validity;
- studies are then excluded if they are not acceptable on the basis of fatal threats to internal and statistical conclusion validity; and
- the remaining studies are coded into two categories of studies (good and bad) on the basis of whether the study design is likely to lead to biased results.

Coding study features. Coding study features represents at least 90% of the work in conducting a meta-analysis (Stock. 1994). Stock has carefully described the following five steps in coding study features: operationally defining all variables, developing a code book. training coders, establishing the reliability of coders. and recording decisions. Stock pointed out that coding is an extremely long process requiring constant vigilance. He also pointed out that it is not unusual for the research questions to change during the process. This necessitates the inclusion of new or different variables and the recoding of all studies.

Pigott (1994) has described three simple methods of dealing with missing data on study features: analyze only complete data sets. substitute the mean. median or some other value for missing data. values, or estimate missing data using regression techniques. She also described two complex methods involving modelling the hypothetical complete data set. However, these complex methods have not as yet been used in meta-analysis. Finally. Pigott offered the practical suggestion of including a code for missing data and testing whether the missing data is correlated to outcomes or other study features.

Studies vary enormously in reporting the quality and completeness of their methodological descriptions. These deficiencies reduce the reliability and validity of coding. Although a complete code book which addresses problem areas and coder training can reduce errors. some will always occur. Domain-specific knowledge can reduce errors in judgement; however. such expertise can also introduce coder bias by reflecting the bias of the coder more than the bias of other experts. Orwin (1994) suggested that after all attempts have been made to reduce coding errors, interrater reliability should be determined. He described four interrater reliability indices: agreement rate or % agreement. Cohen's kappa. the interrater correlation. and the interclass correlation. He stated that all except agreement rate are good indices. Agreement rate poses a problem because when the variables are categorical. chance agreement is often unacceptably high. When variables are ordinal, agreement rate does not discriminate degrees of disagreement. Since the above reliability indexes are not directly comparable, reviewers should report not only the interrater reliability but also the index used. In addition, Orwin (1994) also suggested that researchers need to move beyond describing the coding process to explaining why coders disagree.

In conclusion. Abrami. Cohen. & d'Apollonia (1988) recommended and subsequently used (Abrami, d'Apollonia. & Cohen, 1990) nomological coding to determine the potential study features affecting study outcomes. I recommend that in addition to selecting the study features. the meta-analyst construct a hierarchical representation of these variables. emphasizing the relationships among these variables and between these variables and the outcomes of interest. The meta-analyst should include multiple indicators of study quality and not attempt to derive one global quality index. A code book should be developed and the reliability in coding assessed. In conclusion. the coding of potential explanatory variables should be carried out in a rigorous and systematic manner. Since they form the basis for the rudimentary hypotheses by which the reviewer hopes to explain variability in outcome. the adequacy of the selection of variables and the reliability of coding must be open to assessment.

Data Analysis and Interpretation

Data-analysis refers to the choice of analytical methods (unit of analysis. fixed/random, univariate/multivariate), the calculation of population parameters. the calculation of measures of homogeneity, and the determination of moderator effects (model testing). Data interpretation. on the other hand. includes evaluating the statistical limitations of the analysis. and the threats to internal and external validity. Since data analysis techniques are fully described elsewhere (Cooper. & Hedges. 1994), I will only briefly describe these methods, emphasizing the statistical limitations that hinder interpretation.

Choice of analytical method. Currently, there are a number of well established meta-analytic techniques. Bangert-Drowns (1986) described four approaches to meta-analysis; one which uses the study as the unit of analysis (Kulik. & Kulik. 1988; 1989), one which uses the finding as the unit of analysis (Glass. McGaw. & Smith. 1981) and two which use the pooled-subject as the unit of analysis (Hedges. & Olkin. 1985; Hunter. Schmidt. & Jackson. 1982). These methods differ in complexity; however. software packages. *e.g.*, (Johnson. 1989) are now available for these approaches. In addition to the choice of unit of analysis. decisions need to be made on whether to use a fixed effects or a random effects model and on whether to use univariate or multivariate statistical approaches. These choices are described below:

<u>Unit of analysis</u>. d'Apollonia and Abrami (1988) and Feldman (1989) explored

the consequences of the choice of analytical method in meta-analyses on the multisection validity literature. They found that varying the unit of analysis significantly changed the results of the analyses (see page 137). They recommended that the choice of unit of analysis should be based on the purposes of the review. If the researcher is interested in exploring the influence of study features and statistically determining the homogeneity of subsets of data, he or she must choose the subject as the unit of analysis and use the methods described by Hedges and Olkin (1985) or Hunter and Jackson (1990).

Fixed or random effects models. Shadish and Haddock (1994) reviewed and briefly described both fixed effects and random-effects models of combining estimates of effect sizes. Hedges (1994) and Raudenbush (1994) described fixed and random effects models, respectively, in much greater detail. In the fixed effects models, the population effect size which is estimated from the effect sizes reported in the set of studies is assumed to be fixed at a given value (*e.g.*, for the null hypothesis, it is set at 0). There are only two sources of error which contribute to the differences among the individual effect sizes reported in the studies. One source is the sampling error (non-systematic error) associated with each study, and the other source is the variance (systematic error) explained by study features. The goal of the analyst is to discover the study features which explain the systematic error.

Most meta-analyses have utilized fixed effects models; however, they have not been successful in accounting for the systematic variation in the set of studies (Shadish, & Haddock, 1994). That is, there is significant heterogeneity remaining to "be explained" and under these conditions, significance tests in the fixed effects models are too liberal

(Shadish and Haddock, 1994, p. 275). Thus, most analysts strongly recommend, that if a set of studies is heterogeneous and a fixed effects model is used, the interpretation of significance tests be done cautiously (Hedges, & Olkin, 1985; Gleser, & Olkin, 1994; Matt, & Cook, 1994; Raudenbush, Becker, & Kalaian, 1988).

In the random effects model, the population effect sizes estimated from the effect sizes in the set of studies are not fixed; but rather, are distributed randomly. Therefore, there is a third source of error in the random effects model, *i.e.*, a sampling error (non-systematic error) associated with the population parameter. The random effects model takes this source of error into consideration. Although this model is recommended when a data set is heterogeneous (Becker, & Schram, 1994), it has rarely been used.

Univariate versus multivariate analysis. If the data are independent, univariate analyses are appropriate. However, if the data set contains dependent findings, these dependencies can be modelled by using multivariate analyses (Gleser, & Olkin, 1994; Hedges, & Olkin, 1985; Raudenbush, Becker, & Kalaian, 1988). This approach, like that of Rosenthal and Rubin (1986), takes into consideration the correlations among dependent variables. The multivariate approach differs from that of Rosenthal and Rubin in that it does not advocate the calculation of composite scores but rather explicitly models the interdependencies using a generalized least squares (GLS) regression approach. For independent data, GLS analysis gives identical results to those from a weighted least square regression analysis: however, when data are non-independent (covariances $\neq 0$) the weighted least square analysis is incorrect (Raudenbush, Becker, & Kalaian, 1988, p. 115). The use of univariate methods when the data set is not independent increases the sampling variance about the mean effect size or magnitude (Hunter, & Schmidt, 1990; Strube, 1986). Thus, it does not affect the estimates of central tendency. However, non-independence has a large effect on the homogeneity statistic, inflating the size of Q_T.

Calculation of population parameters. The "heart" of a meta-analysis is the description of what is the overall effect of a given treatment (*e.g.*, the mean effect size of a school library on academic performance) or the overall relationship between two variables (*e.g.*, the mean effect magnitude between student ratings of instruction and student learning). Each study provides one or more estimates of the overall effect across the population of interest. However, individual studies usually vary in sample size and therefore have different sampling variances. In the Glassian approach to meta-analysis, this variation in sampling variance is not taken into account and all studies are treated equally. However, the approaches which use the pooled-subject as the unit of analysis (Hedges, & Olkin, 1985; Hunter, Schmidt, & Jackson, 1982) take the variation in sampling variances into account: studies with smaller sample sizes (and larger sampling variances) contribute less in calculating the average. Thus, in these methods the population mean effect size is a weighted average of all the estimates; where each estimate is weighted by the inverse of its sampling variance.

The approaches which use the pooled-subject as the unit of analysis also allow for the computation of confidence intervals. However, there is controversy on whether the number of outcomes, or the number of subjects should be basis of this computation. Methods based on either Hedges and Olkin (1985) or Hunter and Schmidt (1990) utilize the subject as the unit of analysis. Therefore, the confidence intervals are "dramatically more optimistic" (Rosenthal, 1995, p. 187) than those that use the number of outcomes as the unit of analysis. This is one of the major reservations that Kulik and Kulik (1988, 1989) have with meta-analytic methods employing the subject as the unit of analysis. Rosenthal also suggested using the number of outcomes, and **not** the number of subjects to compute confidence intervals. Combined (Stouffer) Z or other significance tests (see Hedges, & Olkin, 1985) can then be used to determine whether treatment effects differ among each other or from a given value (*e.g.*, 0).

Homogeneity analysis. One of the advantages of selecting the subject rather than the finding as the unit of analysis is that these procedures include a statistic (Q_t) to test the homogeneity of a set of studies. This homogeneity test was developed by Hedges (1982a, 1982b) and is identical to the U test (Marasciulo, 1971). It is the sum of squared deviations about the mean effect size or magnitude divided by the population sampling variance. It has a chi-square distribution with k-1 degrees of freedom. where k is the number of effect sizes. A significant value indicates that the data set includes more than one population.

Although the Type I error rate is acceptable (Rasmussen. & Loher. 1988: Spector. & Levine, 1987), the Type II error rate is variable. depending on the number of studies. the sample sizes per study. and the true population mean effect sizes or magnitudes (Hunter. & Schmidt. 1990; Rasmussen. & Loher. 1988). In general. power is unacceptably low at low sample sizes and unacceptably high at high sample sizes. Therefore. at low sample sizes it fails to indicate that moderators are present: while at high sample sizes it indicates that trivial moderators are present. Since the goal of the meta-analyst is to *fail to reject the null hypothesis*. that is. to achieve homogeneous data sets by controlling for moderator variables, power or Type II error rate is of paramount concern. Because of the variability in Type II error rate. Hunter and Schmidt (1990) recommended that the search for moderators be based on substantiative rather than on statistical criteria.

Determination of moderator effects. Researchers using a homogeneity test based method (Gleser. & Olkin. 1994: Hedges. & Olkin. 1985: Raudenbush. Becker. & Kalaian. 1988) successively break down a heterogeneous data set into subsets on the basis of significant study characteristics until the subset is homogeneous. Thus. these methods rely on significance tests to guide the process. For example. when using the regression methods developed by Hedges and Olkin (1985), meta-analysts would use Q_T to test whether a data-set is homogeneous. If it is not. they would add one or more study features to the model. They would use Q_E to test the goodness of fit of the predictor model, and Q_R to test if one or more predictors are significant[6]. They would continue adding study features until they arrived at a "good" model.

In general. most meta-analysts have not been able to build "good" models of the variability in the studies indicating. that significant variability remains to be explained. The major cause of this problem may be that meta-analysis is not a primary study. The researcher must rely on the study features reported by the primary researchers to explain the variability. This gives rise to a number of problems:

- Power is often too low to detect moderators. As suggested by Hunter and Schmidt (1990). in the absence of a theoretical model. the total sample size (number of studies multiplied by the average sample size per study) would need to be in the order of 2000 to detect moderators. Abrami. Cohen. and d'Apollonia (1988) discussed the importance of considering the power of the significance tests in interpreting the analysis of study features.
- Often primary researchers neglect to describe the study features the meta-analyst

[6] The corresponding statistical tests in the variance-partitioning method are Q_T to test whether the data set is homogeneous, Q_W to test the significance of the within-study and residual variance, and Q_B to test the significance of the between-study or explained variance.

subsequently investigates. Thus, the data set contains a very large amount of missing data. In addition, the investigation of some moderator variables may be hampered by *range-restriction* and unequal sample sizes. For example, in the multisection validity study, there are very few two-semester courses, sophomore, senior or graduate courses, *etc*. If one of these study features were an important moderator such that one level produced a low mean validity coefficient and another level produced a high validity coefficient, the large number of studies with missing data on this variable would insure that a good model could *never* be specified for the complete data set.

The meta-analyst's search for moderators is beset by a number of problems, not faced by the primary researcher. Thus, meta-analysts have recommended that the search for moderators be guided by a theoretical model (Hunter, & Schmidt, 1990; Stock, 1994). This will reduce, but not eliminate these problems, given that the analysis of study features depends on correlational analysis with its interpretation difficulties.

Interpretation. Data do *not* "speak for themselves". Rather researchers must make inferences (based on their analyses of the specific data set) and generalize to the phenomenon in question. The meta-analyst, like other social scientists, should consider potential threats to his or her inferences. Matt and Cook (1994), basing their characterization of the threats to validity on those described by Campbell and Stanley (1966) and Cook, and Campbell (1979), categorized the threats to valid conclusions in a meta-analysis as threats to statistical conclusion validity, to internal validity, and to external validity. Statistical conclusion validity refers to confusion arising from the limitations in data collection and analysis that prevent the meta-analyst from stating accurately whether a relationship exists among the variables of interest. Most of these threats have been described as limitations or problems at the various steps in the meta-analysis, described above (*e.g.*, lack of statistical independence, file-drawer problem, missing studies, *etc.*). In some meta-analytical approaches, attempts are made to control or eliminate these problem; however in many cases the meta-analyst will either not be able to eliminate these threats or choose not to. For example, although the homogeneity approaches (Hedges, & Olkin, 1985; Raudenbush, Becker, & Kalaian, 1988) may have problems with TYPE II error rate, they are still useful techniques, especially for testing models. Thus, in these cases, the meta-analyst must interpret the results with caution, pointing out the statistical limitations.

Threats to internal validity refer to confusion about causal influences that prevent the meta-analyst from stating accurately that one variable (the independent variable) causes the outcome (the dependent variable). Threats to the internal validity of meta-analysis can occur when the threats to internal validity in the primary studies combine across studies to bias the outcome. Slavin (1984) points out that many poorly designed studies suffer from a systematic bias. For example, in the meta-analysis conducted by Carlberg and Kavale (1980) on the effect of student placement on social and academic outcomes, students in regular classes were usually matched on the basis of IQ with students in special education classes. As Slavin points out, students with equivalent IQ's, but placed in different classes (one regular and the other special education) are likely to differ on other characteristics such as having social or behavioral problems. Thus, matching is likely to results in biases in the same direction across all such poorly-designed studies. This systematic bias will threaten the causal inferences made across studies, as well as causal inferences made within studies.

Moreover, in meta-analysis the investigation of the influence of one variable on another is *not* based on an experimental design (true or quasi-): but, rather on a correlational design. Meta-analysis thus suffers from all the limitations of correlational designs, including difficulty in determining causality (Pedhazur, 1982). Moreover, there often is a high degree of collinearity (the predictor variables are highly correlated) in the data set. This multicollinearity produces unstable matrices in the data-analysis . In addition, multicollinearity threatens interpretations of the regression equations, since the same variance can be explained by more than one variable.

There are two types of threats to external validity: threats to generalizability and threats to construct validity (Walberg, 1994). Threats to generalizability refer to confusion concerning the generalizability of the findings that prevent the meta-analyst from stating accurately that his or her inferences apply to research using other subjects, settings, and conditions. Threats to construct validity refer to confusion that arises when the independent (treatment) or dependent (outcome) measures have more than one meaning or operational definition. Thus, these threats prevent the meta-analyst from stating accurately that his or her inferences apply to research using other operational definitions, i.e., measures. This threat is similar to Slavin's (1986) germaneness criteria.

However, as discussed on page 9, meta-analyses are intended to generalize across a variety of subjects, settings, times, and operational definitions. Thus, the question of threats to external validity will always apply to these heterogeneous data sets. The homogeneity-based approaches offer a systematic method of subdividing a heterogeneous data set into homogeneous subsets. Alternatively, postulating and testing models that explain the heterogeneity on the basis of study features, also address threats to external validity. However, since these threats are cumulative (*i.e.*, threats to statistical conclusion cause threats to internal and external validity), the meta-analyst must interpretation the results of these tests with caution.

Implications for Current Research

Meta-analysis is a powerful tool for the integration of findings from a set of similar studies. More than the traditional review, meta-analysis promotes precision and reliability in all aspects of the review process. It applies the general principles of experimental design to the collection of studies, coding of outcomes and study features, data analysis, and hypothesis testing. Other scientists should be able to replicate the findings of any meta-analysis. In addition, meta-analysis, more than an individual study, increases the generalization of findings across diverse populations, settings, and conditions. It does so both by providing estimates of an overall effect size and by exploring the relationships between study features and outcomes. However, as Abrami, Cohen, and d'Apollonia (1988) discussed in their comparison of the six meta-analyses conducted on the validity of student ratings, implementation difficulties abound. The meta-analyst, like any other artisan must use his or her tools wisely, extending their use into new territories but remaining cognisant of their limitations.

In the meta-analyses reported in this thesis, I will search both formally and informally for all studies that meet the inclusion criteria and extract all non-redundant outcomes. I will use nomological coding (Abrami, d'Apollonia, & Cohen, 1990) to develop a coding schema for potential moderators (study features) of the outcomes. Subsequently, I will use multivariate homogeneity approaches (Gleser, & Olkin, 1994; Hedges, & Olkin, 1985; Raudenbush, Becker, & Kalaian, 1988) to model the dependencies within the data set and test a number of models which attempt to explain the discrepancies in reported outcomes.

I will conduct two meta-analyses on multisection validation studies of student ratings of instruction. In the first meta-analysis, I will adapt the above multivariate approaches to the integration of factor studies and address the issue of the dimensionality of student ratings of instruction. Thus, the outcome will not be a scalar common metric (*i.e.*. the mean effect size or magnitude) but rather. a multivariate common metric (i.e.. a mean correlation matrix). I will subsequently use the results of the factor integration to code the outcomes in the second meta-analysis. In this second meta-analysis. I will integrate multisection validity studies to address the issue of the validity of student ratings of instruction and attempt to explain why the studies report such discrepant validity coefficients.

PART II

THE DIMENSIONALITY OF STUDENT RATINGS OF INSTRUCTION

Literature Review

Many post-secondary institutions have adopted the use of student ratings of instruction as one (often the most influential) measure of instructional effectiveness. Student ratings have been used by administrators to help make personnel decisions. by researchers to study the teaching-learning process. by faculty as feedback for self-improvement, and by students for course selection. Student ratings are pencil-and-paper instruments in which students are requested to judge one or more characteristics of their course or instructor by selecting responses on a Likert scale. Typically, student ratings contain global items. such as, *the instructor was excellent,* or. *I would take another course with this instructor*; and/or items assessing specific instructional behaviours (*e.g.*., rapport, course difficulty. feedback. course organization). such as. *the instructor was easy to speak to. the course was too difficult. the instructor never corrected our assignments.* and. *the instructor was well organized.*

The dimensionality. reliability. validity, generalizability. and utility of student ratings have been extensively reviewed (Abrami. d'Apollonia. & Rosenfield. 1996: Costin. Greenough. & Menges. 1971: Feldman. 1978; Kulik. & McKeachie. 1975: Marsh. 1984. 1987; McKeachie. 1979; Murray, 1980). In general. student ratings have been shown to be " clearly multidimensional. quite reliable. reasonably valid. relatively uncontaminated by many variables often seen as sources of potential bias, and are seen to be useful by students, faculty, and administrators." (Marsh, 1987, p. 369). Nevertheless, instructors and researchers continue to express concerns especially when summative evaluation is used for personnel decisions.

Some critics question whether student rating scales are good measures of teaching, *i.e.*, do they measure effective instructional behaviours. Such process-oriented views of instruction (Abrami, d'Apollonia, & Rosenfield, 1996) raise concerns about the structure or dimensionality of student rating forms. Although faculty usually agree that teaching is multidimensional, they often disagree on the interpretation of the interrelationships among instructional dimensions, especially between the specific dimensions and global dimensions. For example, they may view a strong positive correlation between *Instructor Clarity* and *Overall Instructional Effectiveness* as measurement error (*illusory halo*), or as an indicator of the true dependency of specific judgements on global judgements (*true halo*) arising from either general impressions or dimensional similarities (Balzer, & Sulsky, 1992).

Researchers, instructors, and administrators may hold different views of effective instruction, disagreeing on both the goals and means of achieving academic success. They subsequently construct or select student rating forms which reflect their implicit theories of effective instruction, and therefore differ markedly in dimensional structure. Thus, not all student evaluations of instruction are assessing the same instructional behaviours. Although researchers may agree on the dimensionality of a *particular* student rating form, they can and often disagree on the generalizability of its dimensionality to other conditions and to other student rating forms. Thus, one must also address the dimensionality of instructional effectiveness *across* rating forms.

Finally, faculty may disagree on the relationships between ratings (specific, and global) and student learning. This product-process approach to effective instruction emphasizes the causal relationships between specific instructional processes (*e.g.*, providing feedback, clarity, etc.) and learning outcomes or products (*i.e.*, cognitive, affective, psychomotor). Some critics question the validity of student ratings, suggesting that student ratings of instruction are *not* related to student learning. I will address this criticism in PART III, when I review research on the validity of student ratings of instruction. However, a related question, *Should summative evaluations of instruction be based on a single global score or on a set of specific factor scores?*, is related to the question of the dimensionality of instruction and will therefore be discussed here.

I will first review the literature on the dimensionality or structure of individual student rating forms, critically discussing both the dimensions of effective instruction and their interrelationships. I will subsequently review research that has attempted to explore the dimensionality of effective instruction across rating forms. Finally, I will address the controversy surrounding whether summative evaluation should be based on a single score (based on either a global item or on a carefully weighted average of factor scores) or on several specific factor scores.

Factor Studies of Individual Student Rating Forms

There is general agreement that most student rating forms are multidimensional. For example, Marsh (1987) identified nine instructional dimensions[7] on the Students' Evaluation of Educational Quality (SEEQ); Frey, Leonard, and Beatty (1975) identified seven instructional dimensions[8] on the Endeavor; and, Linn, Centra, and Tucker (1975) identified six dimensions[9] on the Students' Instructional Report (SIR). Rating forms can differ both in the instructional behaviours (dimensions) which they assess and in the structure (interrelationships among dimensions) of student ratings. When these rating forms are used to assess effective instruction, they become *de facto* the operational definitions of effective instruction. Thus, the dimensions become the salient indicators of effective instruction; while the factor structure becomes synonymous with the structure of instruction.

Dimensions of Effective Instruction

There are a large number of instructional behaviours which can be used to describe instructional effectiveness. For example, Feldman (1976) identified 19

[7] *Learning/Value, Enthusiasm, Organization, Group Interaction, Individual Rapport, Breadth of Coverage, Examination/Grading,* Assignments, and *Workload/Difficulty.*

[8] *Presentation Clarity, Workload, Personal Attention, Class Discussion, Organization/Planning,* Grading, and *Student Accomplishments.*

[9] T*eacher-student relationship, Course objectives and organization, Course difficulty and workload, Reading assignments, Examinations,* and *Student effort*

dimensions of effective university teachers by reviewing studies which either investigated students' statements about superior teachers or correlated students' evaluation of instructors with instructor characteristics. These dimensions were designed to capture the range of items and factors found on multidimensional rating forms. In subsequent reports, Feldman (1983; 1984;1989) revised this system by adding categories. Feldman's dimensions have frequently been used as a basis of comparison of student rating forms (Abrami, 1988; Abrami, & d'Apollonia, 1990; Feldman, 1989; Marsh, 1991).

Table 1 shows the twenty-eight specific and three global instructional dimensions listed by Feldman (1989) in his meta-analysis of the multisection validity studies. Some dimensions, *e.g.*, *Personality Characteristics ("Personality") of the Teacher* are quite broad; while, others, *e.g.*, *Instructor Pursued and Met Course Objectives*, are quite narrow. Marsh, (1993) considered the Feldman categories to be unidimensional since Feldman had based his reviews on factor studies. However, these rationally derived dimensions vary widely in scope with some being multidimensional and others overlapping (Abrami, & d'Apollonia, 1990). The fact that a researcher can describe a behaviour is no guarantee that the behaviour exists as a distinct construct. I will, therefore, subsequently refer to rationally derived instructional characteristics as *categories* and restrict the term *dimensions* to those derived by factor analysis.

Interpretation of Factor Studies as Evidence of Construct Validity

There have been many factor studies that have investigated the structure of *individual* student rating forms (see p. 84). Perhaps the most studied instructional rating form is the SEEQ, developed by Marsh (1982a). Because of both the quantity and quality of the research on the factor structure of this instrument, I will focus on the research on the dimensionality of this instrument. Factor analyses of the SEEQ (1982a, 1982b, 1983, 1984) collected from over 5000 classes across different courses, disciplines, institutions, and countries have consistently shown that it contains nine correlated factors. Marsh (1982b, 1983a; Marsh, & Hocevar, 1984) demonstrated that the same nine factors are present in peer and self ratings and used the correspondence among factor structures (multitrait-multimethod approach)[10] as evidence for the construct validity of the SEEQ. Marsh (1991a) subsequently used confirmatory factor analysis (CFA) to test one first-order and four alternative second- order structures of the SEEQ. He demonstrated that the nine-factor model was consistent with the design of the instrument. He also demonstrated that one-, two-, three, and four- factor second-order models fit the data with the four-factor second order model having the best fit. However, Marsh recommended that researchers *not* use the second-order factor scores instead of the first-order factor scores to summarize students' evaluation of instruction. He recommended that if specific

[10] The use of the multitrait-multimethod approach to construct validity will be discussed further on page 118..

Table 1. *Thirty-one Instructional Dimensions from Feldman, 1989*

Instructional Category
Teacher's Stimulation of Interests in the Course and Its Subject Matter
Teacher's Enthusiasm (for Subject or for Teaching)
Teacher's Knowledge of the Subject
Teacher's Intellectual Expansiveness (and Intelligence)
Teacher's Preparation: Organization of the Course
Clarity and Understandableness
Teacher's Elocutionary Skills
Teacher's Sensitivity to. and Concern with. Class Level and Progress
Clarity of Course Objectives and Requirements
Nature and Value of the Course Material (Including Its Usefulness and Relevance)
Nature and Usefulness of Supplementary Materials and Teaching Aids
Perceived Outcome or Impact of Instruction
Instructor's Fairness:: Impartiality of Evaluation: Quality of Examinations
Personality Characteristics ("Personality") of the Teacher
Nature. Quality. and Frequency of Feedback from the Teacher to Students
Teacher's Encouragement of Questions and Discussion. and Openness to Opinions of Others
Intellectual Challenge and Encouragement of Independent Thought (by the Teacher and the Course)
Teacher's Concern and Respect for Students: Friendliness of the Teacher
Teacher's Availability and Helpfulness
Teacher Motivates Students to Do Their Best: High Standard of Performance Required
Teacher's Encouragement of Self-Initiated Learning
Teacher's Productivity in Research and Related activities
Difficulty of the Course (and Workload)-Description
Difficulty of the Course (and Workload)-Evaluation
Classroom Management
Pleasantness of Classroom Atmosphere
Individualization of Teaching
Instructor Pursued and Met Course Objectives
Overall Rating of Lectures as an Item of a Multi-item Indicator
Overall Rating of Teacher as an Item of a Multi-item Indicator
Overall Rating of Course as an Item of a Multi-item Indicator

factor scores were not to be used, an appropriately weighted[11] average score or only the overall score could be used for summative decisions. However, Chau (1994) conducted confirmatory factor analysis on the responses to the SEEQ from over 6000 undergraduate classes and demonstrated that three higher order factors were present, corresponding to the three factors described by Widlak, McDaniel, and Feldhusen (1973) and recommended that these second-order factors be used to construct composite scores for summative evaluations.

Abrami and d'Apollonia (1991) commented on Marsh's (1991a) paper and argued that Marsh had misused CFA by using one factor analysis to confirm a second factor analysis rather than using factor analysis to confirm theory. Furthermore, they argued that the use of factor analysis alone, even CFA, to determine the structure of an instrument or construct has the following methodological problems.

- The nature of the items in a rating form strongly influences the resulting factor analysis. For example, if the rating form contains several clusters of similar items, a factor solution consisting of several equally important factors will emerge. On the other hand, if the rating form consists of unique items, a large general factor will emerge.
- The choice of factor analytical methods is based on prior, often unstated, assumptions. This choice subsequently determines the factor solution. For example, factor analysis conducted without rotation is designed to resolve one

[11] This issue will be discussed further on page 66

general or global component. which explains most of the variance. and a few less important, subsidiary factors. Rotation, redistributes the variance explained by the general factor over subsidiary factors and thus, is designed to resolve specific factors of equal importance (Harman. 1976). Thus. care must be exercised in comparing the factor structures obtained by different methods

- The above factor solutions are indeterminate (Gorsuch. 1983). There exist many solutions. each resolving exactly the same amount of the total variance. Therefore, the solution that best describes reality cannot be determined empirically.
- The decision of the number of factors to extract is perhaps the most crucial decision made in factor analysis. Zwick and Velicer (1986) explored the usefulness of six decision rules for extracting the correct number of factors. They reported that Kaiser's rule of only extracting factors that have eigenvalues greater than 1 severely overestimates the correct number of factors. Marsh (1982. 1983a, 1984. 1991a) seriously overextracted factors and built minor factors at the expense of the principle component. Abrami and d'Apollonia (1991) reanalyzed Marsh's data and reported that six. rather than nine, components had eigenvalues greater than 1.0. Therefore. the number of *distinct* specific dimensions may be much less than the number of postulated instructional categories. even for a well designed and validated instrument. like the SEEQ.

Thus. Marsh's factor studies, exploratory or confirmatory, rather than providing evidence for the "true" or "best" structure of effective instruction. reflect the implicit theories held by the rating constructors. When researchers use factor analysis with such circular reasoning they use "statistics as a drunken man - for support rather than for illumination" (Andrew Lang, 1904).

Interpretations of Factor Studies as Evidence for Rater Cognitive Processes

The factor studies of the SEEQ (Marsh, 1982a, 1982b, 1983a, 1983, 1984; Marsh, & Hocevar, 1984), have shown that the nine specific factors are highly correlated. For example, the correlations between the first factor *Learning Value*, and the other eight factors *Enthusiasm, Organization, Group Interaction, Individual Rapport, Breadth, Examination, Assignments*, and *Workload /Difficulty* are .45, .52, .37, .22, .49, .48, .52, and .06, respectively (Marsh, 1984). Since global items (*overall course rating* and *overall instructor rating*) are included in the first two factors, respectively, one might expect that these two factors would correlate highly with other factors. However, it is surprising that *Organization* would be highly correlated with *Breadth* and *Examinations*, *i.e.*, .56 and .57 (Marsh, 1984). The high correlations among supposedly conceptually distinct dimensions and the factorial invariance of student rating forms, such as the SEEQ, has been interpreted by some researchers as reflecting not the true factor structure of student ratings but rather, students' cognitive processes before and during the rating task.

Since Thorndike (1920), high and generally equal correlations among the scales of a rating instrument have been considered to be a symptom of halo effect, defined originally by Thorndike (1920, p. 25) as the "marked tendency to think of the person in general as rather good or rather inferior and to colour the judgements of [specific performance dimensions] by this general feeling" (Balzer. & Sulsky, 1992). Although there has been a massive literature associated with the halo effect, much carried out with student ratings of instruction, there still is confusion and ambiguity concerning conceptual and operational definitions of the halo effect, the sources of the halo effect, and whether halo effects should be treated as error (Balzar, & Sulsky, 1992; Murphy, Jako, & Anhalt, 1993). There are two types of halo; *general impression halo* and *dimensional similarity halo* (Balzer, & Sulsky, 1992). In terms of student ratings of instruction, *general impression halo* is the tendency for students to rate specific dimensions of instruction on the basis of their global evaluation; while *dimensional similarity impression* is the tendency for students to rate similarly the specific dimensions they perceive to be logically or conceptually related. Balzer and Sulskey (1990), surveyed the literature on halo published between 1980 and 1990 and found 108 operational definitions of halo, which they grouped into six categories. The following three methods of operationally defining halo are relevant to the dimensionality of student ratings:

- estimating halo on the basis of average interdimensional correlations across a number of ratees (used in MTMM analysis);
- estimating halo on the basis of the percent of variance accounted for by a single or small number of factors (used in factor analysis); and,
- estimating halo on the basis of partialling out the variance accounted for by the global assessment from that explained by the specific ratings (used in multiple-regression analysis) .

Information-processing theories (Cadwell. & Jenkins. 1985; Cooper. 1981: Jenkins. 1987; Kishor. 1995; Nathan. & Lord. 1983) suggest that cognitive limitations and schematic processing during the impression formation process and rating task give rise to *general impression* (Fiske, & Neuberg, 1990) and *dimensional similarity halo* (Judd. Drake, Downing, & Krosnick, 1991). Individuals minimize cognitive load by using common prototypes to organize their schemas of person. role. and event. These schemas are organized hierarchically, such that general impressions and traits are supraordinate with specific behaviours subordinate. Information-processing models of performance rating contend that such schemas reflect the rater's implicit personality theory of the occupation being rated. Raters may minimize their cognitive load by using supraordinate features. like general impressions, to attend to. store. retrieve. and integrate judgements of specific behaviours. In addition, items on rating forms function as retrieval cues. To the degree that items are semantically similar, they will activate the same node in the rater's associative network (Anderson, 1983), without necessarily involving the raters' general impressions. Furthermore. since persons. roles. and behaviours are stored together. once an individual's implicit personality schema is activated. roles can infer specific behaviours. or *vice-versa* (Trzebinski. 1985). Factors such as the rater's knowledge of the performance domain. familiarity with the instructor. fatigue. item similarity. and the salience of the rating task influence the magnitude and significance of both forms of halo. Thus, factor studies of performance ratings reflect the structure of the performance being assessed, the structure of the raters' implicit theories, and the conditions under which ratings are being collected.

Traditionally, halo has been considered as a rating error, inflating the true correlation among dimensions and reducing the accuracy of rating (Murphy, Jako, & Anhalt, 1993). Thus researchers using performance ratings have sought to eliminate the influence of halo by a variety of means such as statistical control, pooling over raters, training raters, and manipulating the order of items on questionnaires (Cooper, 1981). It should be noted that factor rotation, especially oblique rotation, functions to remove halo by redistributing the shared variance over specific dimensions. However, these methods have been severely criticized (Becker, & Cardy, 1986; Cooper, 1981; Murphy, Jako, & Anhalt, 1993; Nathan, & Tippins, 1990).

They argue that the observed factor structure includes shared variance from a number of sources: from the true correlation among dimensions, from the net result of cognitive processes resulting from dimensional similarity and general impression, and from systematic response- set errors (*e.g.*, range restriction). Some of these sources of halo (true correlation among dimensions, and some of covariation due to cognitive processes) not only do not distort ratings, they may increase accuracy (Murphy, Jako, & Anhalt, 1993; Nathan, & Tippins, 1990). These sources of covariation therefore, represent *true halo*, not *illusory halo* and should not be removed. Moreover, in both laboratory and field settings, Murphy and Anhalt (1992) demonstrated that the ratio of *illusory* to *true halo* varies greatly as a function of rating conditions. Thus, while it may be possible to separate *true* and *illusory halo* in theory, in practice, especially in field settings, this exercise becomes fruitless given current measures of halo (Murphy, & Anhelt, 1992; Murphy, Jako, & Anhelt, 1993).

Whitely and Doyle (1976) demonstrated in a classroom setting that students do have implicit personality theories of teaching (implicit theories of teaching), and that the factor structure of their theories corresponds to the factor structure of their ratings. They also suggested that if students' implicit theories of teaching were idiosyncratic, pooling over students (*i.e.*, using class means) would eliminate this random error. Therefore, they contended that student ratings of instruction were not contaminated by halo error; but rather reflected students' *valid* implicit theories of teaching. Larson (1979) suggested that since these implicit theories were not idiosyncratic, but rather reflected societal norms, their influence on student ratings could not be eliminated by using class means. Thus, he agreed with Whitely and Doyle that factor studies should not be used as evidence for the construct validity of student ratings.

Marsh (1982b, 1983b) used Campbell-Fiske multitrait-multimethod (Campbell, & Fiske, 1959) and ANOVA (Kavanagh, Mackinney, & Wolins, 1971) analyses to assess the extent of a halo effect on the SEEQ. The results indicated that there is a halo effect with student ratings, but not with instructor self-ratings. The ANOVA results suggest that the halo effect is significant ($p < .01$), and accounts for 19% of the variance in student ratings. Whether this is *true* or *illusory halo* cannot be determined: nevertheless, it does indicate a large dependency of specific ratings on an overall assessment of instructional effectiveness.

Cadwell and Jenkins (1985) and Kishor (1995) carried out laboratory studies on the influence of implicit theories of teaching on student ratings of instruction. Cadwell and Jenkins made three assumptions in their study:

- that students hold implicit theories of teaching;
- that semantic similarities among specific behavioral impressions mediate the halo effect; and.
- that under adverse rating conditions, students' ratings of instruction reflect their implicit theories rather than actual instructional behaviours.

They conducted an experiment, using a fractional factorial design (Kirk, 1982), in which graduate students were asked to rate nine hypothetical instructors on 12 SEEQ items loading on *Enthusiasm*, *Organization*, *Group Interaction*, and *Breadth of Coverage*. The hypothetical instructors were described by means of a check list containing eight behavioral descriptors, two for each factor. Instructor profiles were randomly generated such that the check list indicated the presence of two, one, or no behavioral descriptors for each factor. They subsequently factor analyzed the students ratings with and without the influence of the behavioral information statistically removed. Since the two factor analyses were identical, Cadwell and Jenkins (1985) concluded that under conditions of incomplete information, students' implicit theories of teaching had a causal influence on their ratings.

Marsh and Groves (1987) questioned the statistical conclusion validity (inappropriate design, inaccurate degrees of freedom, presence of response set-error), the internal validity (the putative manipulation of implicit theories by manipulating semantic similarities), and the external validity (generalizablity of laboratory study to field settings) of Cadwell and Jenkin's (1985) interpretation. Jenkins (1987) justified the research design, pointing out that Marsh and Groves had forgotten to include the number of hypothetical instructors in their computation of degrees of freedom. and stating that although Marsh and Groves' data simulation may have contained response-set errors, the results of the Cadwell and Jenkins indicated that response-set was not an issue. They justified their manipulation of students' implicit theories by pointing out that students consistently responded to the appropriate cues, as would be expected if they held *valid* implicit theories of teaching.

Cadwell and Jenkins (1985) discussed evidence from cognitive psychology indicating that individuals use both general impressions and semantic similarity as a simplification strategy while rating instruction. They also suggested that the inferences they made on the basis of their laboratory study would be threatened if and only if students in more naturalistic settings were *less* likely to use general impressions or semantic similarities to guide their judgements.

Kishor (1995) used a laboratory study in which he manipulated information on teaching behaviours. similar to that of Cadwell and Jenkins (1985). to test two competing models of how implicit theories of teaching influence student ratings of instruction. The *theoretical model* posited that under conditions of incomplete information. observed behavioral information about an instructor activates the student's person schemas (implicit theories of teaching). which are subsequently used to rate the instructor. The *competing model* posited that under conditions of incomplete information. students will use observed behavioral information about an instructor to both rate the instructor and activate their implied theories of teaching; however, students' ratings would only be based on the observed behavioral information. He used linear structural modelling to demonstrate that the theoretical model best fit the data. and that the hypothesized causal paths were statistically significant. Moreover. the fit of the competing model was significantly worse than that of the theoretical model. Kishor concluded that student ratings of instruction reflect. not only observed behaviours. but also students' implicit theories of teaching. Therefore. he concluded that the robustness of the factor structure of student ratings could not be used as evidence for the structure of teaching, without considering the cognitive processes that underlie rating.

In conclusion, it appears that the factor structure of individual student rating instruments, such as the SEEQ, are robust across a wide selection of courses. institutions. and raters. Some researchers, notably Marsh and his colleagues. have interpreted this invariance as evidence for the construct validity of student ratings of instruction. Therefore, they contend that effective instruction can best be evaluated in terms of multiple (nine), nearly equal factors or traits. However, there appears to be a large general halo-component to performance ratings. including the SEEQ. Other researchers. especially in the area of social psychology. contend that this large halo component reflects human cognitive processes during rating. which may or may not reduce rater accuracy. They contend that the robustness of the factor structure of student rating instruments reflects students (and faculty's) shared implicit theories of teaching.

The Structure of Effective Instruction Across Student Rating Forms

Abrami and d'Apollonia (1990) suggested that Marsh's (1991) conclusions concerning the dimensionality of effective instruction were limited in that he had analyzed one and only one student rating form. Moreover, they suggested that the more general question that needed addressing was the dimensionality of instructional effectiveness ***across*** student rating forms. This question has been approached both logically and empirically. Several researchers (Feldman, 1976; Kulik, & McKeachie, 1975; Marsh, 1987, 1991; Widlak, McDaniel, & Feldhusen, 1973) have attempted to logically integrate factor studies across forms purporting to measure the same dimensions of effective instruction. For example, Widlak, *et al.* (1973) compared twenty-two student rating forms and found that three categories, describing the three instructional roles: actor, interactor, and director could be used to describe the set of factor studies. In all cases, the first factor fell either into the actor or interactor category. They subsequently factor analyzed responses to the 18 item Course-Instructor-Evaluation from Purdue and obtained three highly correlated factors.

Feldman (1976) explored dimensionality across rating forms, using a schema which contained 19 specific and two global instructional categories. He used a clustering technique to show that three interrelated clusters of instructional behaviours exist across all forms. Table 2 lists the three clusters he identified as being associated with three teacher roles: *presentation, facilitation*, and *regulation*. On the basis of the three patterns of intercorrelations, he concluded that his clusters were similar to those described by Widlak, McDaniel, and Feldhusen (1973).

Thus, across student rating forms, teaching is evaluated in terms of three roles. The first role is that of an actor, in which the instructor in command of the domain (both depth and breadth) uses his or her presentation skills (preparation, elocution, sensitivity to audience) to motivate students and to present information in a clear and organized manner. The two global assessments (overall course and overall instructor) are associated with the role of instructor as actor. This may reflect the nature of teaching in most introductory college courses. The second role is that of a facilitator in which the instructor, sets an interactive learning environment by being friendly and tolerant of students' opinions, encouraging students to grow intellectually through both group discussion and independent work, and by being available to students if they need additional help. The third role is that of a manager, in which the instructor manages

Table 2. *Feldman's (1976) second-order "factor structure"compared to that of Widlak et al.(1973).*

Feldman's Instructional Categories	Feldman's Instructional Function	Widlak's Instructional Function
Overall Rating of Instructor Overall Rating of Course Instructor's Stimulation of Interest Instructor's Clarity and Understandableness Instructor's Preparation and Organization Instructor's Enthusiasm Instructor's Knowledge of Subject Outcome of Instruction Instructor's Intellectual Expansiveness Instructor's Elocutionary Skill Instructor's Sensitivity to Class Level and Progress	Presentation	Actor
Instructor's Friendliness, Concern, or Respect for Students Instructor's Openness to Others' Opinions and Encouragement of Class Questions and Discussion Instructor's Intellectual Challenge and Encouragement of Independent Thought Instructor's Availability and Helpfulness	Facilitation	Interactor
Classroom Management Instructor's Fairness, Impartiality of Evaluation and Quality of Evaluation Nature and Value of Course Material Clarity of Course Objectives and Requirements Difficulty of Course and Workload Quality and Frequency of Feedback	Regulation	Director

materials, tasks, and feedback, by insuring that class time is used efficiently and feedback, both formative and summative, is relevant, timely, and fair.

Marsh (1987) also commented on the similarity of specific factors in several student rating forms. Marsh (1991) subsequently compared the correspondence among the SEEQ and Endeavor factors (Frey, Leonard, & Beatty, 1975), and Feldman's (1976) categories. He concluded that Feldman dimensions were much narrower than the SEEQ and Endeavor factors; that the SEEQ represented more of Feldman's categories than did the Endeavor, and that many SEEQ factors contained more than one Feldman category.

Kulik and McKeachie (1975) reviewed eleven studies and identified four commonly found factors which they labelled *Skill*, *Rapport*, *Structure*, and *Difficulty*. Cohen (1981, 1982) explored the validity of student ratings across these four factors and two additional factors from McKeachie, Milholland, Lin, Hofeller, Baerwaldt, and Zinn (1964). However, subsequent research by Cohen (1987, 1988) and d'Apollonia and Abrami (1987, 1988) indicated that even with the addition of other categories (*Motivation*, *Evaluation*, and *Student's Perception of Self-Progress*), many dimensions of effective instruction described in the literature could not be categorized by these nine categories.

In addition to the methodological problems with the factor analysis of individual rating forms described on page 51, Feldman (1976) and Abrami and d'Apollonia (1991) add the following two problems when analyzing the factor structure **across** forms:

- Situational variables, such as the nature of the students responding to the rating forms, the institutions and courses at which ratings are collected, and the rating forms used, influence the factor structure. Although the factor structure of the SEEQ has been shown to be invariant across academic disciplines, instructor rank, and course levels (Marsh, & Hocevar, 1984, 1991), this robustness does not necessarily extend to other rating forms, nor to atypical students or classes. Moreover, as discussed above, this robustness may reflect the stability of implicit theories of teaching rather than that of ratings.

- The factor structure of the rating form depends on whether class means or individual scores are selected as the unit of analysis (Cranton, & Smith, 1990). Thus, care must be exercised in only comparing factor studies that employ the same unit of analysis.

In addition to these two problems, rater errors within individual ratings may not cancel each other out when ratings are aggregated. For example, although the rater errors due to response-set may cancel each other out when individual responses are used in a factor analysis, these errors may not cancel each other out when class mean responses are used in factor analysis (Marsh, & Groves, 1987). This is similar to the issue in meta-analysis where the influence of methodological weaknesses within studies may be predominantly in one direction and bias the estimation of mean effect size or magnitude.

Despite these problems in comparing or integrating different factor studies, researchers (Bushman, Cooper, & Lemke, 1991; Thomson, 1989) have emphasised the need for an empirical approach to integrating factor studies. Kaiser, Hunka, and Bianchini (1971) developed a method by which pairs of factor analyses are compared by projecting the factor and variable axes into a common factor space, rotating the axes to their "best fit" relative to each other, and measuring discrepancies in the variable and factor vector positions. This method has been further developed and used by a number of researchers, both in the area of student ratings of instruction (Rosenfield, d'Apollonia, & Abrami, 1993; Whitely, & Doyle, 1976) and in the general psychology literature (Gorsuch, 1983; Thompson, 1989; Bushman, Cooper, & Lemke, 1991). However, there are two shortcomings of this method. Firstly, it conducts pairwise contrasts rather than comparing all factors simultaneously. Secondly, it relates factors rather than matching (Rosenfield, d'Apollonia, & Abrami, 1993). That is, it correlates specified pairs of factors by computing a "pseudo" correlation between two factors.

Multidimensional Rating Scores and Summative Decisions

Most researchers agree that teaching is multidimensional, and that many student rating forms are multidimensional. However, faculty do not necessarily agree on which instructional dimensions are most important in student learning. There is ample evidence from multisection validity studies (see PART III) that not all dimensions are equally correlated to student learning. For example, there is general agreement that global assessments of teaching (*overall instructor*, *overall course*, and *overall learning*) have uniformly moderately high validity coefficients (Cohen, 1981, 1986; d'Apollonia and Abrami, 1988; Feldman, 1989). However, there is much greater diversity among the validity coefficients for specific dimensions of teaching. Cohen (1986), reported that the validity coefficients ranged from .03 for *Course Difficulty* to .42 for *Skill*. The validity coefficients for the Feldman dimensions are also highly diverse (d'Apollonia and Abrami, 1988; Feldman, 1989) with such dimensions as *Intellectual Challenge, Encouragement of Independent Thought, Nature and Value Course Materials, Nature and Usefulness of Supplementary Materials* and *Course Difficulty* having very low validity coefficients and *Clarity, Stimulation of Interest, and Teacher's Preparation and Organization,* having high validity coefficients. Nevertheless, most post-secondary administrators must, at some point, judge the teaching effectiveness of individual faculty to decide on such issues as retention, merit pay, promotion, and tenure. How are they to interpret student ratings of instruction ?

Some researchers (Frey, Leonard, and Beatty, 1975; Marsh, 1984, 1987, 1989, 1991a, 1991b; Marsh, & Dunkin, 1992) have made strong recommendations that summative evaluations be based on a set of specific scores derived from multidimensional rating forms. Proponents of using a set of factor scores rather than a single score argue that since teaching is multifaceted, it can best be evaluated by a measure or measures that reflect this dimensionality. Since there is little agreement on which pedagogical behaviours necessarily define effective instruction, the specific factors in the set can be differentially weighted to meet individual needs.

Other researchers (Abrami, 1985, 1988, 1989; Abrami, & d'Apollonia, 1990, 1991; Cashin , & Downey, 1992, Cashin, Downey, & Sixbury, 1994) have argued, equally vigorously, that global ratings (or a single score representing a weighted average of the specific ratings) are more appropriate. They argue that the inferences of instructional effectiveness based on specific ratings suffer from threats to internal and external validity and that their use by typical administrators is faulty.

The possibility that student ratings of instruction reflect either their general impressions of the professor or their implicit theories of teaching rather than observed behaviours, may threaten the internal validity of using specific factor scores to infer instructional effectiveness. Whether this constitutes a threat depends on whether the observed halo is *illusory* halo, reflecting rater error, or *true* halo (reflecting the accuracy of students' general impressions and/or implicit theories of teaching). If it is the former, the interpretation of the specific factors as measures of instructional effectiveness is inappropriate. As mentioned previously, MTMM analysis (Marsh, 1982b, 1983a) indicates that even the SEEQ, which is carefully constructed and factor analyzed to produce specific factors, contains a large halo effect. If halo is completely *true* halo, reflecting the students' cognitive strategies before and during rating, the interpretation of specific factors as indicating instructional effectiveness is not necessarily threatened. However, a more immediate indicator of effective instruction would be either a global assessment of teaching effectiveness or the correspondence (a single score) between the students' implicit theory of teaching and the pattern of the instructional behaviours. However, it may be impossible in field settings to distinguish between *true* and *illusory* halo (Murphy, & Anhelt, 1992; Murphy, Jako, & Anhelt, 1993). For this reason, personnel psychologists have begun recommending that global ratings, rather than specific behavioral factors, be used for personnel decisions (Nathan, & Tippins; 1990).

Cashin and his colleagues(Cashin, & Downey, 1992; Cashin, Downley, & Sixbury, 1994) correlated global and specific ratings of effective instruction with a weighted composite score reflecting instructional effectiveness. This criterion measure was a weighted average of 10 IDEA (Instructional Development and Effectiveness Assessment) items rated by students which measure the degree to which course objectives have been met. The weights are assigned by faculty to reflect the importance of the objective to specific courses. They reported that the global items accounted for more than 50% of the variance in the achievement measure. The only specific items that added any practical increment in explanation were *Stimulated students to greater effort*. and *Degree to which the course hung together*. Together these items only added 15%. They made a strong recommendation that global items predict most of the variance in instructional effectiveness and therefore should be used for summative evaluation.

Marsh (1994. 1995) has criticized the Cashin and Downey' (1992) study on the use of global scores for summative evaluation. However, he based his arguments on the use of the weighted composite score as a valid measure of instructional effectiveness, rather than on the appropriateness of using global scores. Moreover. Marsh showed that global scores were highly correlated with his own composite achievement measure.

The recommendation to use factor scores for summative evaluation has been challenged on the basis of threats to external validity: both to generalizability and construct validity (Abrami. & d'Apollonia. 1990). There is little known about the generalization of specific ratings compared to that of global ratings (Abrami. 1989) except for one rating form the SEEQ (Marsh and Hocevar (1984. 1991). For this one rating form, there is evidence that the factor structure (*i.e.*. the number and nature of the teaching dimensions found) and thus the relationships among perceived characteristics of teaching was stable across those variables that were studied. However, this interpretation of the factor invariance of the SEEQ has been questioned. As mentioned previously, the factor invariance of student rating forms can be interpreted as reflecting students' information processing before and during the rating task. rather than the construct validity of a multidimensional interpretation of student ratings.

Abrami and d'Apollonia (1990) explored the threat to the construct validity of using specific factor scores to infer instructional effectiveness by determining the consistency and uniformity in dimensional structure exhibited by multisection validity studies. They coded the items from the multidimensional student rating forms contributing to each reported validity coefficient into 21 specific. two global. and one miscellaneous dimension using a coding schema derived from that of Feldman (1976, 1983, 1984). They reasoned that if there were a more or less consistent dimensional structure across student rating forms. the dimensions would contribute equally across findings. However. the most frequent dimension (*Clarity and Understandableness*) contributed to more than 67% of the findings: while the least frequent (*Instructor's Enthusiasm*) contributed to less than 20% of the findings. Moreover. they reasoned. that each dimension would be found in an equally "pure form" across findings (uniformity index = 1). That is. dimensions would not be unidimensional in some forms but multidimensional in others. However. the uniformity index varied from a low of .11 for *Intellectual Expansiveness* to a high of .61 for *Overall Instructor*. That is. multidimensional rating forms are inconsistent in their treatment of items denoting *Intellectual Expansiveness*, but consistent in their treatment of *Overall Instructor*. Thus, using specific factor scores to infer instructional effectiveness does not generalize across different operational definitions or constructs of effective instruction.

Franklin and Theall (1990) have suggested that administrators. untrained in instructional evaluation, may not have the necessary expertise to properly weigh the numerous factor scores and arrive at a single decision. Administrators appear to simply average over all the instructional dimensions. Since specific ratings are neither equally valid indices of student learning, nor equally relevant across different courses (Abrami. & d'Apollonia, 1990; Abrami, d'Apollonia, & Rosenfield. 1996). it is inappropriate to make comparative judgements based on such summative scores.

Implications of Literature for This Thesis

The lack of consensus on the dimensional structure of effective instruction as projected by student ratings of instruction poses problems for both theoretical and applied research. Researchers must develop a circumscribing model or "common metric" with which to compare the different structures. That is. they must construct a model of teaching effectiveness that encompasses more than one student rating form. The degree to which global scores correlate with specific scores should also be addressed. That is. an attempt should be made to determine whether student rating forms in general reflect a general teaching factor or multiple. more or less equal. specific factors. This common structure can subsequently be used to explore other characteristics of student rating forms. for example their validity. Thus. one of the goals of this thesis was to adapt meta-analytical methods to the aggregation of the factor studies of student ratings of instruction. This integration will allow me to determine whether there is a common structure of teaching effectiveness (as perceived by students) across forms, and whether it supports the contention that students judge teaching on the basis of a general teaching factor or multiple specific factors.

Method

This section describes the application of meta-analytical techniques to the determination of the common structure of student rating forms. The following four steps will be used as the framework for the description of the methods used for the integration of factor studies: problem formulation and specification of inclusion criteria, identification of studies, coding items, and data analysis.

Problem Formulation and Specification of Inclusion Criteria

The problem investigated in the meta-analysis of the factor studies of student ratings of instruction was, "What is the structure of instructional effectiveness (as judged by students) across student rating forms?". The following inclusion criteria were used.

- Since the results of this meta-analysis were intended to provide a "common-metric" for the meta-analysis of the multisection validity studies, factor studies had to be conducted on the rating forms used in these studies.
- The complete factor matrix or interitem correlation matrix had to be available. When more than one factor or correlation matrix was available the one most closely resembling the rating form used in the multisection validity studies was used. If more than one was similar, the matrix was chosen randomly.

Identification of Studies

The studies reporting the factor structure or interitem correlation matrix of the student rating form used in the validity studies were identified. These were located by branching from the references in the original primary studies and by conducting a computerized search on ERIC using the authors' names and the name of the student rating form in question as key words. These studies are described on page 84

Coding Outcomes

One of the problems that the reviewer of this literature faces is the diversity of instruments used by the primary researchers. Some primary researchers report the validity coefficients for single items while other researchers report the validities of factors (either empirically or rationally derived). Since the rating forms are not uniform, it is necessary to code the correlation coefficients in terms of the items they represent. Thus, this section describes the development of the coding scheme that was subsequently used to code the items into categories.

Development of Coding Schema

The coding schema was based on those developed by Feldman (1976, 1983, 1984). However, since he did not supply operational definitions for the categories, but rather gave exemplars, the operational definitions had to be inferred from the exemplars. To do so, the following method was used. I listed all the examples Feldman had used and sorted the items by category. After clarifying the ambiguities among the exemplars within each dimension, I developed a provisional set of operational definitions. Two research assistants then coded all the items in the student rating forms used by Feldman. The overall inter-rater reliability was 0.93 (Cohen's *kappa*). Abrami and d'Apollonia (1990) subsequently used this schema to investigate the dimensionality and uniformity across the student rating forms used in the multisection validity studies.

Feldman further modified his system and coded a sub-set of the rating forms used in the multisection validity literature (Feldman, 1990). This system consists of 28 specific dimensions and three global categories. To further characterize the reliability of the coding schema, I compared the inter-rater reliability between our categorization (Abrami. & d'Apollonia. 1990) and Feldman's (1989) for the common studies. The inter-rater reliability (Cohen's alpha) was only 0.60.

There are at least two reasons for this lower reliability. one trivial and the other substantiative. First. I used the item as the unit of analysis. while Feldman used the number of dimensions per validity coefficient (factors) as his unit of analysis. and did not weight the dimensions by the number of items. In order to compute the inter-rater reliability between our categorization and Feldman's. I collapsed on factors and eliminated weighting by items. However. this has the effect of amplifying discrepancies. For example. consider the situation where Feldman and I were coding the same primary study in which data was reported for a factor consisting of ten items. If Feldman were to code the ten items into two categories (eight items into *Enthusiasm for Subject* and two items into *Enthusiasm for Teaching*), he would enter the finding twice, weighting it .50 in each category. However. if I were to disagree with Feldman on only one item coded by

Feldman into *Enthusiasm for Teaching* and by me into *Ability to Stimulate Interest*, I would enter the finding three times, weighting it .1, .8, and .1 in the three categories *Enthusiasm for Teaching, Enthusiasm for Subject* and *Ability to Stimulate Interest*, respectively. Thus, the inter-rater reliability would be .81 when the unit of analysis is the item, but only .65 when the unit of analysis is the dimension with no weighting. Thus, the lower reliability may be an artifact of the choice of unit of analysis.

The more substantiative issue is the lack of operational definitions provided by Feldman, the ambiguity, multidimensionality, and overlap of some of the dimensions, and the ambiguity of some of the items. These difficulties were reported in Abrami and d'Apollonia (1990). For example, in many cases items referring to the instructors behaviour were placed in one category; but, items referring to the results of the behaviour were put in another category.

As a result of the above low inter-rater reliability and the internal inconsistencies, I further modified the coding schema along the following principles:

- The classification schema should not be ambiguous. The categories used to code items should be clear, comprehensive and succinct. At any one level, the categories should be of more or less equal breadth.
- The bipolar values of the categories should be in the same category: *e.g.*, clear and unclear presentations, authoritarian and participatory class management, etc.
- Both the product and the process orientations to a teaching behaviour should be in the same category; for example, the instructor presenting the subject as interesting and the students being interested in the subject.

- Since global evaluations (course instructor, perceived learning) are included, the remaining categories should only include specific statements.

I subsequently developed provisional operational definitions for the 42 categories which were used to code the items in the rating forms in the multisection validity studies.

Coding Items into Categories

I and a research assistant subsequently coded 1190 items extracted from the student rating forms used in the multisection validity studies. The overall interrater reliability (as percent agreement) across categories was 91.5%. Table 3 illustrates the interrater reliabilities for each category. The interrater reliabilities for five categories *(the instructor's research productivity and reputation, the instructor's ability to deliver relevant instruction, the instructor's ability to monitor the class' response, the instructor's supervision and disciplinary actions,* and *the instructor's ability to foster tolerance of diversity*) are below .70. In order to resolve disagreements, all disagreements were investigated. For some categories, *the instructor's ability to deliver clear instruction, relevance of instruction, the instructor's supervision and disciplinary actions* and *management style*, the coders stated that they had simply made clerical errors. However, other disagreements were due to difficulties with applying the coding schema to certain categories. For example, coders found it difficult to decide whether some items referred to the instructor's willingness to respond to students' difficulties during class time (*the instructor's ability to monitor the class' response*) or outside of class time (*availability*). These difficulties were dealt with by clarifying the operational definitions.

Table 3. *Interrater reliability in coding items into instructional categories*

DIMENSION	N of ITEMS	N in AGREEMENT	% AGREEMENT
(AA) Personal Appearance	9	9	100
(AB) Personality Characteristics	18	15	83
(AC) General Attitudes	7	6	86
(BA) Knowledge of Domain	21	21	100
(BB) Knowledge of Teaching	4	4	100
(BC) General Knowledge	5	5	100
(CA) Enthusiasm for Subject	12	12	100
(CB) Enthusiasm for Teaching	13	13	100
(CC) Enthusiasm for Students	18	15	83
(D) Research Productivity	3	2	67
(EA) Choice of Required	32	30	94
(EB) Choice of Supplementary	4	4	100
(F) Preparation and Organization	32	30	94
(GA) Ability to Stimulate Interest	55	49	89
(GB) Ability to Motivate to	28	25	89
(HA) Formulation of Objectives	25	28	89
(HB) Implementation of	14	14	100
(IA) Use of Teaching Methods	36	31	86
(IB) Clarity of Instruction	63	58	92
(IC) Relevance of Instruction	28	18	64
(ID) Response to Questions	17	15	88
(JA) Monitoring Student's	16	16	100
(JB) Monitoring Class' Response	6	2	33
(KA) Vocal Delivery	11	11	100
(KB) Dramatic Delivery	15	15	100
(LA) Management Style	63	55	87
(LB) Time Management	20	20	100
(LC) Discipline	10	15	67
(MA) Interaction	66	59	89
(MB) Tolerance of Diversity	29	19	66
(MC) Respect for Others	90	89	99
(MD) Friendly Classroom	17	15	88
(NA) Low-level Cognitions	25	20	80
(NB) High-level Cognitions	76	68	90
(OA) Concern for Students	15	14	93
(OB) Availability	13	13	100
(P) Feedback	35	33	94
(Q) Workload	53	50	94
(R) Evaluation	60	59	98
(S) Overall Course	36	33	92
(T) Overall Instructor	61	59	97
(U) Overall Learning	29	29	100

Coders also experienced difficulties in distinguishing between certain pairs of categories. It was difficult to distinguish between *monitoring student's response* and *monitoring the class' response*. Often the only difference in the items was whether the apostrophe was before or after the final *s* in the word *students*. Therefore, these categories were merged . Similarly, it was difficult to determine whether an item referred to the instructors's *formulation of* or *implementation of objectives*. Thus, these categories were also merged producing a final coding schema of 40, not 42 categories.

Five items belonging to *the instructor's ability to foster tolerance of diversity* were miscoded by Coder 1 into *management style* . Three additional items were miscoded into categories *classroom interactions* and *respect for others*. These items were difficult to code because the instructor's management style affects the social climate in the classroom, namely *classroom interactions*, *tolerance for diversity*, and *respect for others*. The items were:

- The instructor tries to force us to accept his ideas and interpretations.
- Teacher rejected the students' statements.
- Permitted students to express opinions which differed from his own.
- The instructor invited criticism of his acts.
- The instructor's efforts improved the ability of the students to understand deviant individuals.
- Students argued with each other or with the instructor, not necessarily with hostility.
- Did your instructor remain open to criticism.

- Intolerant

Thus. there is some overlap amongst the categories which is difficult to resolve. The coding schema was modified to take these difficulties into account and is presented in Appendix 1. It contains forty instructional categories that are *assumed* to be unidimensional and logically distinct. The acceptable inter-rater reliability indicates that coders can reliably distinguish among categories. However. it *neither* indicates that students can discriminate among the categories *nor* that they represent first-order factors of instructional effectiveness. These latter questions are addressed by the results of the research integration and will be addressed in the Results and Discussion sections.

Data Analysis

The data analysis of a meta-analysis of factor studies is somewhat complicated because the summary statistics which are being aggregated. the interitem correlation coefficients of the correlation matrix. are multivariate. Broadly speaking. the analysis that was carried out consisted of extracting and selecting a reliable set of interitem correlation coefficients. aggregating them (and computing population parameters) to produce an aggregate correlation matrix. and subsequently factor analyzing this correlation matrix in order to determine the common structure across rating forms.

Extraction of Outcomes

The outcome variables of interest for the integration of factor studies are the interitem Pearson product moment coefficients in each student rating form. These were estimated from the reproduced correlation matrices[12] computed from the factor loading matrix of the items in the student rating forms (if rotated orthogonally), or from the pattern matrix and factor correlation matrix (if rotated obliquely). If only the correlation matrix for a rating form was available, the correlation matrix was factor analyzed to obtain the reproduced correlation matrix. The following formulae were used to compute the reproduced correlation matrices:

$$R = AA'$$

where **R** is the reproduced correlation matrix,
A is the factor loading matrix (rotated orthogonally),
A^t is the transposed factor loading matrix.

$$R = A\phi A'$$

where Φ is the factor correlation matrix.

Both factor loading and pattern matrices are item by factor matrices. When the items measured a negative aspect of instruction, the factor loadings for these items were multiplied by -1. In addition, since the sign of factor loadings is arbitrary (Gorsuch, 1983, p. 181), the factor loadings for any factor with a preponderance of negative loadings was also multiplied by -1.

Selecting Interitem Correlation Coefficients

In order to aggregate the interitem correlation coefficients, one must first establish

[12] The reproduced correlation matrix is the difference between the original interitem correlation matrix and the residual correlation matrix. Thus in a "good" factor structure which extracts most of the item variance, the reproduced correlation matrix will be very similar to the original interitem correlation matrix. It would be identical if all factors were extracted.

that the values being aggregated are homogeneous. If the set is not homogeneous. the weighted mean correlation does not represent each study well.

There are a number of possible causes for heterogeneity:

- the items are ambiguous and/or multidimensional;
- the categories are ambiguous and/or multidimensional; and
- the relationship between items varies with setting, subject. etc.

The first two reasons speak to technical problems with the coding schema and certain student rating forms. Unfortunately, these problems confound questions concerning the dimensionality of effective instruction. Therefore. if one wishes to address the latter question. one must first reduce these technical problems. We therefore, pruned items and categories which were heterogeneous in the following manner.

I and a colleague pruned the items and categories from the data set in a two stage process. In the first stage we eliminated items that contributed to *poor* correlations between items belonging to the same category. We subdivided the complete set of interitem correlations (21,383 correlations) into the forty sets of interitem correlation coefficients between items belonging to the same category. We assumed that. if the categories were unidimensional and generalizable. sets of interitem correlation coefficients should be uniform across student rating forms and the mean interitem correlation coefficient for the set should approach 1.0. In other words. the mean interitem correlation coefficient for the subset is analogous to a reliability coefficient. For each set, we identified items which contributed to correlations which were below 0.5, or which lowered the mean interitem correlation coefficient for a category consistently below .60. We scrutinized these items for ambiguous wording, reversed polarity, negative wording, compound items, etc. We subsequently dropped these items. For each set, we continued pruning until the set was homogeneous (i.e., the coefficient of variability was .20 or less).

In the second stage we eliminated items that contributed to heterogeneous correlations between items in different categories. This is a more difficult task in that it can not be assumed that the correlations should approach 1.0. However, they should cluster about the weighted mean. We subsequently subdivided the remaining correlations into sets representing the intercorrelations between items belonging to different categories. Taking one set at a time, we identified the items that contributed to correlations at the extremes of the distribution. We eliminated those items that consistently contributed to the heterogeneity of the set. Finally, after all pruning had been done, we reviewed all decisions to see if later decisions to drop items would allow us to reinsert some dropped items. Note that two items contribute to a correlation that is an *outlier*. In some cases the *poor* item can be easily identified because of poor wording, double negatives, compound items, etc.. However, in other cases which item is to be eliminated is somewhat arbitrary since it may be a poor item by virtue of its position in the rating form. In other words, there is *not* one unique set of items which, if eliminated, produces homogeneous sets. Rather, there are a number of possible sets.

Synthesis of Aggregate Correlation Matrix

The interitem correlation coefficients and categories (35) that survived selection were subdivided into sets representing the cells of the aggregate correlation matrix (a 35x35 matrix). The following procedures adapted from Hedges and Olkin (1989) were used to compute the population parameters. The mean interitem correlation coefficient was computed using the following formula:

$$z_{+} = \sum_{i=1}^{k} w_i z_i$$

where z_+ is the weighted mean interitem correlation coefficient.
z_i is the Fisher z transformation of each correlation.
w_i is the weight for each study based on the sampling variance calculated from the following formula.

$$w_i = \frac{n_i - 3}{\sum_{i=1}^{k} (n_i - 3)}$$

n_i is the sample size in each study.

The coefficient of variation was used as a measure of the degree of heterogeneity in the set. This measure is suggested by Hunter and Schmidt (1990) since the power of the Q_T is very high when study sample sizes and/or the number of studies is high. In the factor studies. the sample size can be as high as 5000. In addition. Q_T . gives the probability that the observed variance is greater than would be expected on the basis of sampling error. It does not give any indication of the degree of heterogeneity (Rosenthal. 1995). We used the following formulae to calculate the corrected sampling variance:

$$\sigma_p^2 = \sigma_r^2 - \sigma_e^2$$

where σ_p^2 is the corrected sampling variance.
σ_r^2 is the observed variance among the interitem correlations in the sample. calculated as described below,
σ_e^2 is the sampling variance for the set of interitem correlation coefficients calculated as described below.

The observed variance for the set of interitem correlation coefficients is calculated by the following formula:

$$\sigma_r^2 = \frac{\sum_{i=1}^{k} n_i(r_i - r)^2}{\sum_{i=1}^{k} n_i}$$

where **r_+** is the weighted mean interitem correlation coefficient back transformed from **z.**

r_i is the ith interitem correlation coefficient.

n_i is the sample size of the ith interitem correlation coefficient.

The sampling variance for the set of interitem correlation coefficients is calculated by the following formula:

$$\sigma_e^2 = \frac{\sum_{i=1}^{k} \frac{n_i(1 - r_i^2)^2}{n_i - 1}}{\sum_{i=1}^{k} n_i}$$

where **r_+** is the weighted mean interitem correlation coefficient.

r_i is the ith interitem correlation coefficient.

n_i is the sample size of the ith interitem correlation coefficient.

Factor Analyses

First-order factors were extracted from the synthesized correlation matrix using principal components extraction (SPSS Inc.. 1994). Factors with eigenvalues greater than 1.0 were retained. The solution was then rotated both orthogonally using VARIMAX. and obliquely using OBLIMIN. Different values of *delta* were used. The solution that gave the best discrimination among factors (*delta* = 0.4) was subsequently interpreted. Subsequently, the factor matrix from the oblique solution was factor analyzed to extract second-order factors and the solution was rotated orthogonally.

Results

I will describe the factor studies that were included in the meta-analysis and the characteristics of the aggregated correlation matrix at various stages of its synthesis before presenting the results of the primary and secondary factor analyses. These results, the uncovering of the common structure of student ratings of instruction across forms, will subsequently be used as a common framework for the meta-analysis of the multisection validity studies in PART III.

Description of Factor Studies

There are 38 student rating forms used in the 43 primary studies in the multisection validity set (see PART III). Seventeen factor studies of these rating forms were obtained and are listed in Table 4. These seventeen studies produced eight factor matrices and ten correlation matrices. One of the rating forms, the CEQ, provided correlation matrices of two subscales. Thus, almost 50% of the rating forms used in the multi-section validity studies were included in the meta-analysis of the factor studies[13]. However, one student rating form, that used by Wherry (1951), supplied almost 50% of the items in the data set and therefore, supplied a major portion of the interitem correlation coefficients which were subsequently aggregated. A comparison of the distribution of categories in the factor set to the distribution of categories in the student rating forms used in the multisection validity studies is presented in Table 5. As can be

[13] I used two data sets in this thesis. The larger data set consists of the student rating forms in the 43 primary studies in the multisection validity studies. The smaller data set, is a subset of student rating forms for which I could obtain factor or correlation matrices.

seen the global items are under-represented in the factor set relative to the multisection validity studies. Other under-represented categories are *use of teaching materials* (IA), and *interaction* (MA). On the other hand, *personal appearance* (AA) and *personality characteristics* (AB) are over-represented. These items came primarily from the Wherry (1951) student rating form. However, the distribution of categories in the rating forms is not different in the two sets of studies ($X^2 = 32.32$, df = 39).

Table 4. *Characteristics of the factor studies in the data set.*

TRF Name	Factor Study	Matrix Type	n_i	n_f	n_r
Purdue	Bendig (1954)	C	10	3	3
SIR	Linn. Centra. & Tucker (1975)	F	29	6	4
TCE	Bolton (1990)	F	19	5	1
SIRS	Michigan State University (1971)	F	21	5	1
SOS	Doyle. & Crichton (1978)	C	7	2	2
CEQ	Gilmore (1973)	C	8	3	1
CEQ	Gilmore (1973)	C	10	5	1
in house	Endo. & Della-Piana (1976)	C	7	4	1
Endevor	Frey. Leonard. & Beatty (1975)	F	21	7	3
H&H	Hartle. & Hogan (1972)	F	26	10	1
CLIC	Hoffman (1978)	F	12	2	1
SEEQ	Marsh. & Hocevar (1984)	F	35	9	2
SOST	Mintzes (1977)	C	20	5	1
in house	Murdock (1969)	C	5	3	1
MSU	Pambookian (1972)	C	23	6	2
in house	Rankin. Greenmun. & Tracy (1965)	C	4	2	1
in house	Solomon. Roseberg. & Bezdek (1964)	C	7	3	1
in house	Wherry (1951)	F	194	12	1

note C is a correlation matrix. F is a factor structure matrix. n_i is the number of items in the rating form. n_f is the number of factors in the rating form. and n_r is the number of times the rating form appears in the multisection validity literature.

Table 5. *Distribution of items in categories in validity set, factor set and individual student rating forms. The student rating form numbers refer to the numbers given in Table 3.*

CAT	VALIDITY SET		FACTOR SET		STUDENT RATING FORM																	
					1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
	n	%	n	%	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
AA	9	.8	11	2.4	1																	10
AB	18	1.5	24	5.2	3		1			1						1						18
AC	7	.6	4	.9	1																	3
BA	21	1.8	4	.9						1												3
BB	4	.3	1	.2																		1
BC	5	.4	2	.4																		2
CA	12	1.0	3	.7	1			1														1
CB	13	1.1	5	1.1				1	1	1						1						1
CC	18	1.5	11	2.4	1		1			1						1		1				6
D	3	.3	3	.7																		3
EA	32	2.7	9	2.0		1					2	1				2	1					2
EB	4	.3	4	.9		1	1				1											1
F	32	2.7	13	2.8		1		1			2	1	1				2	1				4
GA	55	4.6	21	4.6		1	1	1				1		4		2	2	1		1	2	5
GB	28	2.4	19	4.1	1	1	2	2	1					1			2		4		1	4
H	39	3.3	11	2.4		3		1					1				1		1			3
IA	36	3.0	2	.4		2																0
IB	63	5.3	25	5.5	1	1	2	2		1		1	2			4	2		1			8
IC	28	2.4	11	2.4		1		1					1			4						4
ID	17	1.4	9	2.0			1	1		1		1				1	1					4
J	22	1.8	7	1.5		1	1												1			5

Table 5 cont. *Distribution of items in categories in validity set, factor set and individual student rating forms. Student rating form numbers refer to the numbers given in Table 3.*

CAT	VALIDITY		FACTOR		STUDENT RATING FORM NUMBER																	
	SET		SET		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
	n	%	n	%	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
KA	11	.9	7	1.5			1										1					5
KB	15	1.3	5	1.1												1						4
LA	63	5.3	23	5.0						1			1						2			19
LB	20	1.7	12	2.6		2	1	1			2					1			1			4
LC	10	.8	3	.7																		3
MA	66	5.5	15	3.3		1		1			1		3			3			1			5
MB	29	2.4	12	2.6		2	1	2						1					2			4
MC	90	7.6	28	6.1													1		1			26
MD	17	1.4	8	1.7												1			1			6
NA	25	2.1	9	2.0									1	5	2						1	0
NB	76	6.4	32	7.0		2				1			2	15	9	1					1	1
OA	15	1.3	8	1.7		1		1					2				1					3
OB	13	1.1	7	1.5		1	1						1			2	1					1
P	35	2.9	16	3.5		1	1									1	2		3			7
Q	53	4.5	20	4.4		3	1	3					3			3	2		3			2
R	60	5.0	27	5.9	1	3	2				2	1	3			2	1					12
S	36	3.0	8	1.7				1	1							1			1	3		1
T	61	5.1	11	2.4					3			1				1		1	1		2	2
U	29	2.4	8	1.7			2	1	1						1	1		1				0
TOT	1190		458		10	29	19	21	7	8	10	7	21	26	12	35	20	5	23	4	7	194

Characteristics of Aggregate Correlation Matrix

The seventeen student rating forms in the factor set yielded 458 items, categorized into 40 categories, which furnished 21,383 interitem correlation coefficients. The average correlation, range, standard deviation, number of correlation coefficients, and coefficient of variability for the forty sets of interitem correlations between items belonging to the same category are presented in Table 6. Interitem correlations between items belonging to the same category are only present when student rating forms contain multiple items belonging to the same category; *e.g.*, two items pertaining to an instructor's preparation and organization will yield one interitem correlation; while four items will yield six correlations. There were no student rating forms in the factor set that contained multiple items relevant to *Knowledge of Teaching*, *Enthusiasm for Subject*, *Enthusiasm for Teaching*, and *Choice of Supplementary Materials*. Thus, we have no measure of the reliability of these categories. There are also a large number of categories which have low reliabilities; *e.g.*, the 25 intercorrelations within *Feedback* have an average correlation of only .25. On the other hand, the 156 intercorrelations within Personality Characteristics have a respectable average correlation of .67.

The items were scrutinized and 130 and 65 items were dropped at the first and second pruning stages, respectively. The items pruned at each stage are presented in Appendix 2. Five categories (and their items) were dropped from further analysis because of missing values *(Use of Teaching Materials* (IA), L*ow-level Cognitions* (NA), and O*verall Learning* (U)) or excessive heterogeneity which could not be eliminated (*Time Management* (LB) and W*orkload* (Q)).

Table 6. *The mean reliability (M), range, standard deviation (SD), number of interitem correlations (K), and coefficient of variability (C) for the complete dataset.*

CATEGORY	M	RANGE		SD	K	C
(AA) Personal Appearance	.52	-.11	.91	.28	45	.61
(AB) Personality Characteristics	.67	-.05	.96	.24	156	.40
(AC) General Attitudes	.63	.59	.69	.01	3	.01
(BA) Knowledge of Domain	.13	.02	.28	.09	3	.67
(BB) Knowledge of Teaching					0	
(BC) General Knowledge	.67				1	
(CA) Enthusiasm for Subject					0	
(CB) Enthusiasm for Teaching					0	
(CC) Enthusiasm for Students	.55	.28	.80	.15	15	.28
(D) Research Productivity	.44	.22	.72	.21	3	.51
(EA) Choice of Required Assignments	.37	.19	.87	.08	4	.21
(EB) Choice of Supplementary Materials					0	
(F) Preparation and Organization	.81	.36	.91	.24	8	.33
(GA) Ability to Stimulate Interest	.65	.23	.97	.21	17	.35
(GB) Ability to Motivate to Greater Effort	.46	-.16	.97	.44	15	1.49
(H) Use of Objectives					0	
(IA) Use of Teaching Methods	.36				1	
(IB) Clarity of Instruction	.65	.34	.94	.18	38	.30
(IC) Relevance of Instruction	.58	.12	.91	.28	12	.58
(ID) Response to Questions	.46	.01	.81	.32	6	.82
(J) Monitoring Response to Instruction	.71	.42	.95	.15	10	.22
(KA) Vocal Delivery	.58	.43	.72	.12	10	.21
(KB) Dramatic Delivery	.36	.23	.65	.13	5	.36
(LA) Management Style	.30	-.20	.91	.32	191	1.22
(LB) Time Management	-.09	-.20	.64	.21	7	7
(LC) Discipline	-.06	-.26	.25	.21	3	3.7
(MA) Interaction	.74	.17	.96	.23	6	.34
(MB) Tolerance of Diversity	.58	-.03	.73	.17	9	.30
(MC) Respect for Others	.51	-.12	.96	.20	326	.43
(MD) Friendly Classroom Environment	.67	.42	.88	.15	15	.24
(NA) Low-level Cognitions	.38	.24	.55	.08	11	.20
(NB) High-level Cognitions	.41	.16	.74	.09	143	.22
(OA) Concern for Students	.73	.57	.85	.09	4	.12
(OB) Availability	.83				1	
(P) Feedback	.25	-.11	.71	.20	25	.82
(Q) Workload	.47	.16	.89	.28	14	.70
(R) Evaluation	.40	-.19	.88	.24	64	.66
(S) Overall Course	.41	.35	.46	.18	3	.45
(T) Overall Instructor	.66	.42	.90	.19	5	.32
(U) Overall Learning	.73				1	

The remaining 6788 interitem correlations computed for 227 items from seventeen student rating forms were aggregated to produce the 35 by 35 correlation matrix used in subsequent analyses. The average correlation, range, standard deviation, number of correlation coefficients, and coefficient of variability for these thirty-five sets of interitem correlations between items belonging to the same category are presented in Table 7. The reliability for all categories for which there is data is now greater than 0.60. Figure 1 shows the distribution of the coefficients of variation for the sets of interitem correlations between items in different categories before (a) and after (b) pruning. Before pruning (Figure 1a) there are 665 unique sets of off-diagonal interitem correlation coefficients. Many are extremely heterogeneous, indicating that some of the items may be unreliable and must be dropped. After pruning (Figure 1b) there are 660 sets of unique sets of off-diagonal interitem correlation coefficients. The distribution is now negatively skewed indicating that most sets are now homogeneous. Less than 7% of the sets of interitem correlations have coefficients of variation greater than 0.20. Many of these contain items belonging to *Preparation and Organization* (F), *Monitoring Learning* (J), *Tolerance to Diversity* (MB), *High-Level Cognitions* (NB), *Concern for Students* (OA) and *Feedback* (P). Since additional pruning resulted in the elimination of many categories, no additional items were eliminated.

The aggregated correlation matrix is presented in Table 8. As can be seen many of the correlations between categories are high. There are very few low correlations. Thus, one would expect a highly correlated structure to emerge from the factor analysis.

Table 7. *The mean reliability (M), range, standard deviation (SD), number of interitem correlations (K), and coefficient of variability (C) for the pruned dataset.*

CATEGORY	M	RANGE		SD	K	C
(AA) Personal Appearance	.87	.81	.91	.03	6	.03
(AB)Personality	.78	.52	.96	.10	106	.12
(AC) General Attitudes	.84				1	.00
(BA) Knowledge of Domain					0	
(BB) Knowledge of Teaching					0	
(BC) General Knowledge	.81				1	.00
(CA) Enthusiasm for Subject					0	
(CB) Enthusiasm for Teaching					0	
(CC) Enthusiasm for Students	.69	.63	.80	.05	6	.07
(D) Research Productivity	.90				1	.00
(EA) Choice of Required Assignments	1.00				1	.00
EB) Choice of Supplementary Materials					0	
(F) Preparation and Organization	.82	.53	.92	.16	4	.20
(GA) Ability to Stimulate Interest	.87	.64	.97	.12	12	.14
(GB) Ability to Motivate Greater Effort	.92	.76	.97	.06	7	.07
(H) Use of Objectives					0	
(IB) Clarity	.87	.62	.94	.05	13	.06
(IC) Relevance of Instruction	.84	.72	.92	.07	6	.08
(ID) Response to Questions	1.00				1	.00
(J) Monitoring Learning	.79	.54	.95	.12	6	.15
(KA) Vocal Delivery	.60	.55	.65	.02	3	.03
(KB) Dramatic Delivery	.77				1	.00
(LA) Management Style	.79	.58	.90	.09	22	.11
(LC) Discipline						
(MA) Interaction	.80	.60	.95	.11	7	.13
(MB) Tolerance of Diversity	.69	.65	.72	.02	3	.03
(MC) Respect for Others	.77	.51	.96	.10	79	.13
(MD) Friendly Classroom Environment	.81	.69	.88	.06	6	.07
(NB) High-level Cognitions	.65	.58	.74	.07	4	.10
(OA) Concern for Students	.96				1	.00
(OB) Availability	1.00				1	.00
(P) Feedback						
(R) Evaluation	.78	.66	.89	.10	4	.13
(S) Overall Course						
(T) Overall Instructor	.82	.64	.90	.13	3	0.16

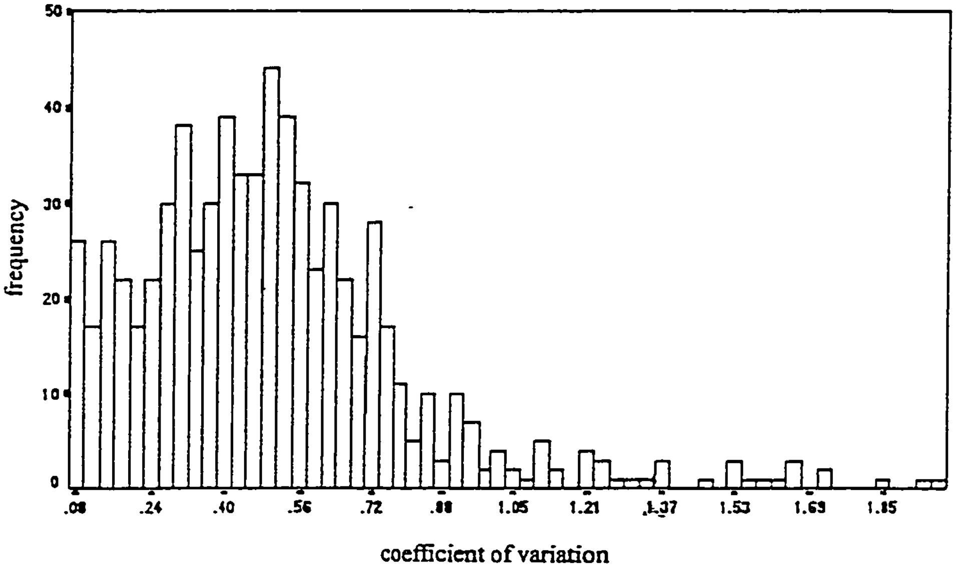


Figure 1a. Distribution of 665 coefficients of variation for the off-diagonal intercorrelation coefficients of the complete data set.

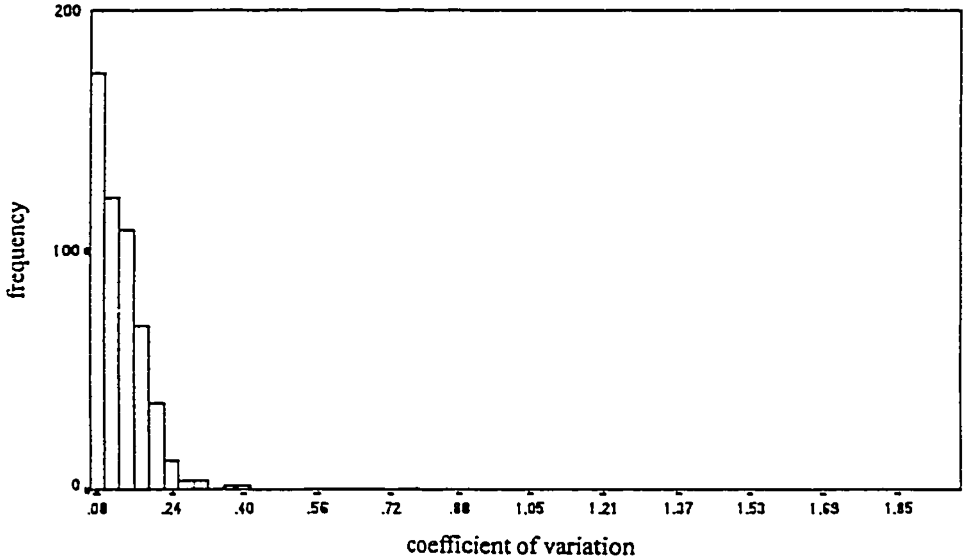


Figure 1b. Distribution of the 600 coefficients of variation for off-diagonal intercorrelation coefficients of the pruned data set.

Table 8. *The aggregate correlation matrix across seventeen student rating forms. Decimal values have been omitted*

	AA	AB	AC	BA	BB	BC	CA	CB	CC	D	EA	EB	F	GA	GB	H	IB	IC	ID	J	KA	KB	LA	LC	MA	MB	MC	MD	NB	OA	OB	P	R	S	T
AA		60	67	24	73	50	49	66	64	61	45	33	48	66	59	29	56	56	70	52	62	71	59	24	65	56	62	65	67	73	65	32	48	46	72
AB	60		61	31	82	67	56	79	71	72	55	55	65	75	81	38	78	70	75	71	60	66	76	41	74	71	75	78	69	77	74	54	57	78	82
AC	67	61		25	67	50	44	62	60	59	45	39	48	60	60	19	55	54	65	54	56	65	58	26	60	61	60	64	60	68	65	30	45	58	66
BA	24	31	25		26	33	29	35	28	33	15	32	30	34	32	13	34	31	33	27	23	32	28	07	29	37	31	35	32	34	38	21	27	35	37
BB	73	82	67	26		70	65	87	77	70	70	50	75	81	85	43	85	69	89	72	75	73	83	57	84	74	79	79	78	83	74	53	59	73	87
BC	50	67	50	33	70		54	82	64	68	48	62	66	75	76	38	75	73	67	67	54	58	70	30	67	63	65	67	65	67	65	45	46	78	76
CA	49	56	44	29	65	54		64	50	52	42	40	57	61	60	29	59	61	63	49	55	55	54	31	59	42	55	53	58	58	52	32	45	54	62
CB	66	79	62	35	87	82	64		78	83	53	70	77	88	86	41	79	72	73	81	67	76	84	36	78	73	79	79	75	80	67	55	61	86	81
CC	64	71	60	28	77	64	50	78		72	53	54	61	73	76	35	71	66	71	70	58	64	73	36	74	65	71	74	75	73	77	53	54	73	79
D	61	72	59	33	70	68	52	83	72		48	65	61	77	79	33	74	78	70	74	54	65	72	24	72	66	70	75	76	71	76	50	50	81	79
EA	45	55	45	15	70	48	42	53	53	48		35	45	55	59	29	58	50	58	51	50	48	58	44	54	52	55	52	54	55	45	47	45	58	59
EB	33	55	39	32	50	62	40	70	54	65	35		52	64	67	27	63	68	48	65	38	48	59	13	55	51	52	58	63	49	60	42	33	74	62
F	48	65	48	30	75	66	57	77	61	61	45	52		71	71	36	72	63	65	61	57	59	67	42	67	64	63	61	66	57	61	42	47	68	70
GA	66	75	60	34	81	75	61	88	73	77	55	64	71		83	40	80	74	75	75	65	71	78	34	76	70	74	74	78	78	67	52	56	82	84
GB	59	81	60	32	85	76	60	86	76	79	59	67	71	83		42	88	82	78	82	61	66	84	42	81	73	78	82	67	78	79	62	55	91	83
H	29	38	19	13	43	38	29	41	35	33	29	27	36	40	42		48	33	36	34	30	28	42	25	38	27	38	39	36	34	36	33	31	40	46
IB	56	78	55	34	85	75	59	79	71	74	58	63	72	80	88	48		76	79	76	60	65	82	45	76	72	77	79	59	65	70	63	45	84	85
IC	56	70	54	31	69	73	61	72	66	78	50	68	63	74	82	33	76		64	78	54	62	76	27	64	45	70	67	55	54	56	54	49	60	77
ID	70	75	65	33	89	67	63	73	71	70	58	48	65	75	78	36	79	64		65	63	72	75	46	78	74	76	77	62	68	70	51	63	72	79

Table 8 cont. *The aggregate correlation matrix across seventeen student rating forms. Decimal points have been omitted*

	AA	AB	AC	BA	BB	BC	CA	CB	CC	D	EA	EB	F	GA	GB	H	IB	IC	ID	J	KA	KB	LA	LC	MA	MB	MC	MD	NB	OA	OB	P	R	S	T
J	52	71	54	27	72	67	49	81	70	74	51	65	61	75	82	34	76	78	65		53	58	74	33	73	63	67	72	78	67	74	58	43	83	77
KA	62	60	56	23	75	54	55	67	58	54	50	38	57	65	61	30	60	54	63	53		62	60	40	64	57	57	55	67	57	59	31	41	51	67
KB	71	66	65	32	73	58	55	76	64	65	48	48	59	71	66	28	65	62	72	58	62		63	27	67	65	65	66	69	72	63	40	56	64	75
LA	59	76	58	28	83	70	54	84	73	72	58	59	67	78	84	42	82	76	75	74	60	63		43	76	68	74	77	73	75	75	57	52	79	82
LC	24	41	26	07	57	30	31	36	36	24	44	13	42	34	42	25	45	27	46	33	40	27	43		42	37	38	35	37	40	35	35	24	31	41
MA	65	74	60	29	84	67	59	78	74	72	54	55	67	76	81	38	76	64	78	73	64	67	76	42		70	73	75	74	77	69	54	54	73	81
MB	56	71	61	37	74	63	42	73	65	66	52	51	64	70	73	27	72	45	74	63	57	65	68	37	70		69	71	55	52	74	41	56	52	77
MC	62	75	60	31	79	65	55	79	71	70	55	52	63	74	78	38	77	70	76	67	57	65	74	38	73	69		77	66	74	73	50	60	75	81
MD	65	78	64	35	79	67	53	79	74	75	52	58	61	74	82	39	79	67	77	72	55	66	77	35	75	71	77		67	78	80	57	61	79	83
NB	67	69	60	32	78	65	58	75	75	76	54	63	66	78	67	36	59	55	62	78	67	69	73	37	74	55	66	67		42	63	40	41	73	79
OA	73	77	68	34	83	67	58	80	73	71	55	49	57	78	78	34	65	54	68	67	57	72	75	40	77	52	74	78	42		79	40	40	56	86
OB	65	74	65	38	74	65	52	67	77	76	45	60	61	67	79	36	70	56	70	74	59	63	75	35	69	74	73	80	63	79		58	58	65	78
P	32	54	30	21	53	45	32	55	53	50	47	42	42	52	62	33	63	54	51	58	31	40	57	35	54	41	50	57	40	40	58		62	63	59
R	48	57	45	27	59	46	45	61	54	50	45	33	47	56	55	31	45	49	63	43	41	56	52	24	54	56	60	61	41	40	58	62		65	65
S	46	78	58	35	73	78	54	86	73	81	58	74	68	82	91	40	84	60	72	83	51	64	79	31	73	52	75	79	73	56	65	63	65		85
T	72	82	66	37	87	76	62	81	79	79	59	62	70	84	83	46	85	77	79	77	67	75	82	41	81	77	81	83	79	86	78	59	65	85	

The Kaiser-Olkin test of sampling adequacy was .82. indicating that the correlation matrix was suitable for factor analysis (SPSS Inc.. 1994). Another indication that the correlation matrix is suitable for factoring is that only 14.8% of the offdiagonal elements in the residual correlation matrix have values > .05 or < -.05.

Primary Factor Analysis of Aggregate Correlation Matrix

Four factors were extracted with principal components extraction from the aggregate correlation matrix using SPSS (SPSS Inc., 1994). The percent variances extracted by each factor, in decreasing magnitude were, 62.8%. 4.2%. 3.7%. and 2.9%. Table 9, the sorted principle components solution, illustrates that of the 35 categories, all except U*se of Course Objectives* (H), *Discipline* (LC), and *Knowledge of Domain* (BA) had high loadings (> 0.60) on the first factor. Thus, there clearly is a large general principal component (PCI) which explains about 63% of the variance in instructional effectiveness across the seventeen student rating forms in the factor set. The correlations between *Overall Instructor* and *Overall Course* and this General Teaching Component (PCI) are 0.94 and 0.88, respectively.

In order to try to get meaningful specific factors. the principal components solution was rotated obliquely using OBLIMIN (SPSS Inc.. 1994) and a delta of .40. Table10. the sorted pattern matrix. indicates that most categories clearly load on only one factor. indicating that the factor solution can be interpreted. Exceptions are, *Overall Instructor*, *Enthusiasm for Subject*, *Personality*, *Availability*, *Interaction*, *Respect for Others*, *Enthusiasm for Students*, and *Use of Objectives.*

Table 9. *The sorted principal components for the 35 categories of instructor effectiveness*

CATEGORY	FACTORS			
	1	2	3	4
(D) Research Productivity	0.95	-0.24	-0.09	-0.05
(T) Overall Instructor	0.94	0.02	-0.03	0.03
(BB) Knowledge of Teaching	0.93	0.24	0.16	-0.03
(GB) Ability to Motivate Greater Effort	0.93	-0.16	0.07	.00
(CB) Enthusiasm for Teaching	0.93	-0.09	-0.06	-0.09
(GA) Ability to Stimulate Interest	0.90	-0.05	-0.06	-0.12
(IB) Clarity	0.89	-0.16	0.18	0.03
(LA) Management Style	0.89	-0.06	0.11	-0.04
(S) Overall Course	0.88	-0.32	0.03	.00
(MD) Friendly Classroom Environment	0.88	-0.02	-0.06	0.12
(AB) Personality	0.88	0.02	0.02	0.05
(MA) Interaction	0.87	0.09	0.04	-0.03
(ID) Response to Questions	0.86	0.21	0.04	0.14
(MC) Respect for Others	0.85	0.05	.00	0.10
(OB) Availability	0.84	0.02	-0.12	0.13
(CC) Enthusiasm for Students	0.84	0.02	-0.03	-0.07
(J) Monitoring Learning	0.84	-0.25	0.01	-0.18
(F) Preparation and Organization	0.83	-0.06	0.12	0.02
(OA) Concern for Students	0.83	0.23	-0.10	.00
(NB) High-level Cognition	0.81	0.01	-0.06	-0.29
(BC) General Knowledge	0.81	-0.20	-0.04	-0.05
(IC) Relevance of Instruction	0.81	-0.23	-0.05	-0.13
(MB) Tolerance of Diversity	0.81	0.06	-0.07	0.19
(KB) Dramatic Delivery	0.79	0.23	-0.23	-0.07
(AA) Personal Appearance	0.73	0.42	-0.25	-0.11
(KA) Vocal Delivery	0.71	0.33	-0.18	-0.19
(AC) General Attitudes	0.71	0.31	-0.04	-0.10
(EB) Choice of Supplementary Materials	0.69	-0.49	0.33	-0.16
(CA) Enthusiasm for Subject	0.68	0.14	0.33	0.05
(EA) Choice of Required Materials	0.64	0.16	-0.09	-0.12
(P) Feedback	0.62	-0.30	0.32	0.13
(R) Evaluation	0.61	0.16	-0.09	0.47
(H) Use of Objectives	0.46	-0.10	0.32	0.28
(LC) Discipline	0.45	0.28	0.64	0.06
(BA) Knowledge of Domain	0.38	-0.15	-0.40	0.49

Table 10. *The sorted factor structure matrix for the 35 categories of instructor effectiveness*

CATEGORY	FACTORS			
	1	2	3	4
(EB) Choice of Supplementary Material	1.00	-0.17	-0.21	-0.04
(IC) Relevance of Instruction	0.79	0.07	0.02	0.02
(J) Monitoring Learning	0.78	0.05	0.06	0.08
(S) Overall Course	0.78	-0.05	0.03	0.28
(BC) General Knowledge	0.75	0.11	0.05	0.01
(D) Research Productivity	0.66	0.27	-0.12	0.10
(CB) Enthusiasm for Teaching	0.65	0.30	0.06	0.08
(GB) Ability to Motivate Greater Effort	0.65	0.15	0.14	0.21
(NB) Higher-order Cognitions	0.64	0.36	0.11	-0.19
(IB) Clarity	0.61	0.09	0.24	0.20
(GA) Ability to Stimulate Interest	0.61	0.32	0.07	0.05
(F) Preparation and Organization	0.56	0.20	0.23	-0.02
(LA) Management Style	0.54	0.22	0.21	0.17
(T) Overall Instructor	0.43	0.40	0.10	0.24
(CA) Enthusiasm for Subject	0.38	0.38	0.16	-0.10
(AB) Personality	0.38	0.35	0.13	0.24
(AA) Personal Appearance	-0.04	0.89	0.00	0.03
(AC) General Attitudes	0.03	0.79	-0.07	0.07
(KB) Dramatic Delivery	0.18	0.68	-0.04	0.09
(OA) Concern for Students	0.18	0.66	0.08	0.11
(KA) Vocal Delivery	0.20	0.63	0.28	-0.18
(BB) Knowledge of Teaching	0.25	0.53	0.37	0.12
(ID) Response to Questions	0.14	0.53	0.20	0.29
(MB) Tolerance of Diversity	0.12	0.51	0.03	0.34
(OB) Availability	0.24	0.44	-0.04	0.39
(MA) Interaction	0.38	0.41	0.19	0.14
(MC) Respect for Others	0.31	0.39	0.11	0.28
(CC) Enthusiasm for Students	0.37	0.38	0.08	0.21
(LC) Discipline	-0.09	0.13	0.78	0.10
(EA) Choice of Required Materials	0.12	0.24	0.45	0.21
(BA) Knowledge of Domain	0.23	0.22	-0.43	0.27
(H) Use of Objectives	0.31	-0.15	0.36	0.19
(R) Evaluation	-0.14	0.26	0.00	0.80
(P) Feedback	0.24	-0.26	0.19	0.76
(MD) Friendly Classroom Environment	0.34	0.36	0.01	0.37

The tentative identification of the factors is described below.

Sixteen categories have their highest loadings on factor 1. These are, in order of factor loading, *Choice ofSupplementary Materials* (EB), *Relevance of Instruction* (IC), *Monitoring Learning* (J), *Overall Course* (S), *General Knowledge* (BC).*Research Productivity* (D*)*, *Enthusiasm for Teaching (CB), Ability to Motivate Greater Effort (GB)*, *Higher-order Cognitions* (NB), *Clarity* (IB),*Ability to Stimulate Interest* (GA*)*, *Preparation and Organization* (F), *Management Style* (LA), *Overall Instructor* (T*)*, *Enthusiasm for Instruction* (CA), *Personality* (AB*)*. The sum of the squared loadings is 8.0; and therefore, this first factor is the most important factor in instructional effectiveness as perceived by students. Because most of these factors refer to the instructor's role in delivering information, I have tentatively identified this factor as similar to that described by Widlak, McDaniel, and Feldhusen (1973) and Feldman (1976) pertaining to the instructor's presentation of material.

Twelve categories have their highest loading on factor 2. These include, *Personal Appearance*(AA), *General Attitudes* (AC), *Dramatic Delivery* (KB), *Concern for Students* (OA), *Vocal Delivery* (KA), *Knowledge ofTeaching* (BB), *Response to Questions* (ID), *Tolerance ofDiversity* (MB), *Availability* (OB), *Interaction* (MA), *Respect for Others* (MC), and *Enthusiasm for Students* (CC). The sum of the squared loadings is 5.6; and therefore, the second factor is almost as important as the first factor. Because most of the categories refer to the instructor's role in facilitating a social learning environment, I have tentatively identified this factor as similar to that described by Widlak, McDaniel, and Feldhusen (1973) and Feldman (1976) pertaining to the instructor's interaction with students.

Four categories have their highest loadings on factor 3. These are *Discipline* (LC), *Choice of Required Materials* (EA), *Knowledge of Domain* (BA) and. *Use of Objectives* (H). The sum of the squared loadings is 1.1 and therefore the third factor is less important than the other two factors. Not only is it a heterogeneous category; it also contains items which only load moderately on this factor.

Three categories have their highest loadings on factor 4. These include, *Evaluation* (R), *Feedback* (P), and *Friendly Classroom Environment* (MD). The sum of the squared loadings is 1.4 and therefore factor 4 is less important than the first two factors, but more important than factor 3. The first two, highly loading, categories refer to the instructor's role in evaluating learning.

The factor correlation matrix, presented in Table 11, indicates that three of the factors are highly correlated. These are factors 1, 2, and 4. Thus, the factor correlation matrix also appears to indicate that there are three highly correlated factors and one miscellaneous factor. These three correlated factors describe the instructor in terms of presenting material, facilitating interaction, and evaluating performance.

Table 11. *The factor correlation matrix*

	Presentation (F1)	Interaction (F2)	Miscellaneou (F3)	Evaluation (F4)
Presentation (F1)	1.00	0.63	0.31	0.56
Interaction (F2)	0.63	1.00	0.28	0.44
Miscellaneous (F3)	0.31	0.28	1.00	0.24
Evaluation (F4)	0.56	0.44	0.24	1.00

Secondary Factor Analyses of Aggregate Correlation Matrix

Because the above factor analysis resulted in correlated factors. a secondary factor analysis of the factor correlation matrix was carried out via principal components extraction. The percent variances extracted by each factor, in decreasing magnitude were, 56.7%, 20.6%, 14.2%, and 8.5%. The eigen values for these four factors were 2.27, .83, .57, and .34, respectively. The first two factors, representing 77% of the variance were rotated orthogonally via varimax rotation. and the results are presented in Table 12 . Factors 1, 2, and 4 load highly (> .79) on the first second-order factor (HIER I), while Factor 3, the least important and most heterogeneous factor, loaded highly on the second second-order factor (HIER II). Thus, the factor analysis suggests that there is a large general hierarchical factor representing the instructor's role in delivering facilitating, and evaluating instruction, and a smaller miscellaneous factor. This factor is very similar to the general teaching factor extracted by principal component extraction (Table 9). The only difference being that the first-order general teaching component (PCI) includes *Choice of Required Materials* and the second-order general teaching factor (HIER I) does not. This first factor (HIER I), has an internal consistency of 0.78 and measures general teaching.

Table 12. *The second order factor structure matrix for the four first-order factors, rotated via varimax*

	HIER (I)	HIER (II)
Presentation (F1)	0.87	0.18
Interaction (F2)	0.80	0.17
Miscellaneous (F3)	0.16	0.98
Evaluation (F4)	0.79	0.07

Discussion

The goal of this meta-analysis was to determine the dimensionality or structure of effective instruction *across* the student rating forms used in the multisection validity studies. The results reported in this thesis indicate that, although student rating forms can include 35 specific instructional behaviours, the forms (at least those used in the multisection validity studies) tend to measure general instructional skill. General instructional skill is a composite of three correlated factors, delivering instruction, facilitating interactions, and evaluating learning. I will discuss the limitations of factor analysis, specifically the limitations of factor analysis across forms and subjects, the structure of effective instruction as perceived by students, and the implications of these results for summative evaluation.

Limitations of Factor Analysis

All the limitations of primary factor analyses discussed on page 51 also apply to the secondary factor analysis carried out in this thesis. More specifically, the selection of 35 "distinct" categories rather than multiple entries for each category may have influenced the final factor solution. The factor solution reported here is only one of many possible solutions. However, I also conducted factor analyses in which I included all 40 categories, in which I excluded data from Wherry (1951), in which I extracted 2 to 6 factors, and in which I rotated the initial factor solution at different delta values. In all cases I obtained similar final factor solutions. The factor solution reported here was the solution that best met the criteria for a *simple structure* described by Gorsuch (1983).

The secondary factor analysis. *i.e.*. the factor analysis of a composite correlation matrix derived from different subjects and different student rating forms gives rise to additional limitations. In conducting the meta-analysis of the factor studies, I synthesized a *common* correlation matrix (more accurately a reproduced correlation matrix) across seventeen student rating forms. Some forms contributed to all cells of the 35x35 matrix, while other forms contributed to only a few cells. Therefore. different studies, and therefore different subjects were used to calculate the different entries in the correlation matrix. This *missing data* problem can. and probably does. produce bizarre and potentially unstable matrices (Gorsuch. 1983).

Marsh (1994) discussed this limitation of the secondary factor analysis. For example, he suggested that it would be preferable to use the covariance matrix. rather than the correlation matrix, to use results from studies using the same student rating form, and if that cannot be done, to use the results from studies using common items. However, Marsh appears to have missed the point that this is a *secondary*. not a *primary* factor analysis. Although it would be preferable if each rating form had the same number of items distributed in the same categories. such uniformity does not exist (Abrami. d'Apollonia. & Cohen. 1990). A meta-analyst attempts to reflect the body of literature as it exists. not as it should be. Moreover. if student rating forms were uniform. a meta-analysis would probably not be necessary. Given that aggregating across forms and subjects is the only way. in a meta-analysis. of analyzing the data. this limitation is unavoidable. Thus, this limitation poses a potential threat to statistical conclusion validity and the impact of this threat must be taken into consideration in the interpretation of the analyses.

A second limitation arises concerning the representativeness of the aggregated correlation matrix. Some student rating forms (*e.g.*, Wherry, 1951) contributed more correlations to the common matrix than did short rating forms (*e.g.*, Rankin, Greenmum, & Tracy, 1965). However, the distribution of the categories in the factor studies does not differ from that in the multisection validity set. Therefore, it does represent the population of interest. Furthermore, analyses were conducted with and without the data from Wherry (1951) and gave similar results.

A third limitation, also mentioned by Marsh (1994), is the poor reliability of some of the items in the student rating forms. Some items are ambiguous, poorly worded, or not easily categorized. In order to obtain a stable correlation matrix, poor items and categories were eliminated from the aggregation. Two analyses were conducted, one with the complete data set, and a second with the pruned data set. The two analyses gave very similar factor solutions. Thus, pruning poor items did not bias the factor solution although it did decrease the standard error about the mean intercategory correlation coefficients. The types of items that were frequently unreliable are discussed below since they have important implications for the construction of student rating forms.

An examination (Appendix 2) of the items that were eliminated indicates that in some cases, the *poor* item can be easily identified because of poor wording, double negatives, compound items. Some items assessed two characteristics simultaneously. For example, *in this class I feel free to question or express my opinion*, or *course material was unorganized and hindered understanding* . Such items were coded (and weighted appropriately) into two categories. However. students rating their instructors, are required to give only one response and therefore such items may reduce the reliability of students' responses.

Some items required that coders and/or students make an inference about the cause of a behaviour. For example, *called on students alphabetically*, or *became confused in class*. Does the observation that the instructor called on students alphabetically indicate that the instructor is *effective* because he or she is organized or does it mean that the instructor is *ineffective*. because he or she is not responding to the students' needs but rather teaching "mechanically" ? There were similar difficulties with items that appear dated or highly subjective. For example, items such as *nice looking*, *a typical "old-maid" or "bachelor" personality* did not correlate highly with other items in the same instructional category.

Items which were bipolar, that is items in which the response indicating effective instruction in the middle of the scale, presented special problems. When responses were not bipolar. high ratings denoted effective instruction: however. when the responses were bipolar. both high and low ratings denoted ineffective instruction. Many of the items categorizing *workload* and *time management* were scaled from *too easy* to *too difficult* and therefore, aggregating across these categories was not possible.

Items with negative intercorrelation coefficients posed a special difficulty in that some negative correlations could not be made consistently positive (by multiplying the intercategory correlation coefficient to which they contributed by -1.0). These negative correlations may reflect clerical errors, and/or students' confusion with negative items. In some cases, two negative items when correlated with each other remain negative because the items are negative in different ways. For example, *the pace of the course was too slow* and *ideas and concepts were delivered too rapidly.* These would both be negatively correlated with positive items such as *excellent course.* Reverse coding would fix the sign of the correlations with the positive item but not with the other negative item.

In some cases it was difficult to perceive why one item would be eliminated and a very similar item retained. There are two reasons for this difficulty. Firstly, items could only be eliminated if two or more items belonging to the same category were present in the same student rating form. Thus, some "unique" poor items were never eliminated. Secondly, the position of an item within the rating form could make it unreliable. For example, some rating forms change the direction of the scale within the questionnaire. Items on either side of the change may correlate poorly even if they belong to the same instructional category.

There are also some indications that some excluded items came predominantly from one rating form. For example, three items assessing *overall instructor* which were pruned came from the SOS (Doyle, & Crighton, 1978). Similarly, many of the excluded items from the *personal appearance and attire* and *management style* categories came from the Wherry (1951) study. In addition, correlations from SIR (Linn, Centra, & Turner, 1975) were often the lowest correlations in a set. These observations may suggest that setting differences may also be involved in the low reliability (and high heterogeneity). In conclusion, there are a number of limitations in attempting to aggregate the responses of different students to different rating forms across different institutions and times. However. if we are to compare the instructional effectiveness of faculty across different courses. departments. institutions. *etc.* there must be a generalizable construct. instructional effectiveness which all student rating forms presumably measure. Barring a large-scale primary study across institutions and courses, meta-analysis with all its shortcomings, provides a method of discerning the properties of this generalizable conception of instructional effectiveness. Thus. in this factor analysis of the aggregated intercategory correlation matrix. evidence was provided that a stable hierarchical factor, *general instructional skill,* underlies student rating forms.

The Structure of Effective Instruction As Perceived by Students

The factor analyses conducted on the aggregated reproduced correlation matrices from seventeen student rating forms indicate that students rate instructors in terms of general instructional skill. General instructional skill is a composite of three correlated factors delivering instruction, facilitating interactions, and evaluating learning. The first factor. represents the instructor's role in delivering information. It includes the global scales (overall course and instructor) as well as such behaviours as monitoring learning, enthusiasm for teaching. clarity, presentation and organization. The second factor, represents the instructor's role in facilitating a social learning environment. It includes such behaviours as general attitudes. concern for students. availability, respect for others, and tolerance of diversity. The third factor. represents the instructor's role in evaluating learning. It includes such behaviours as evaluation, feedback and providing a friendly atmosphere. This factor structure is very similar to that described by Widlak, McDaniel, and Feldhusen (1973) in their examination of twenty-two factor studies.

The above factor structure is also very similar to the three clusters of instruction described by Feldman (1976) in his review of multidimensional student rating forms (see Table 2). If Factor 3 and Factor 4 from this study are combined, eighteen of the twenty instructional categories common to this study and the Feldman (1976) study have the same distribution (compare Table 2 with Table 10). The exceptions are:

- *Instructor's Elocutionary Skill* which in the Feldman study is in the *Presentation* cluster, but is in the *Facilitation* cluster in this study.
- *Instructor's Intellectual Challenge and Encouragement of Independent Thought* which in the Feldman study is in the *Facilitation* cluster, but is in the *Presentation* cluster in this study.

Chau (1994) conducted confirmatory factor analysis on the responses to the SEEQ and also demonstrated the presence of the same three higher order factors. Marsh (1991a) argued that although confirmatory factor analysis of the SEEQ demonstrated the presence of a second-order factor analysis, student ratings should still be interpreted on the basis of the first-order structure. Even the SEEQ, considered by Marsh (1987) to be the best exemplar of a well-constructed multidimensional student rating form measuring distinct specific instructional dimensions, reflects the same underlying factor structure "exposed" by the meta-analysis conducted in this thesis.

In his critique of Cashin and Downey (1992), Marsh (1995) suggests that the objectives in the Instructional Development and Effectiveness Assessment (IDEA) have no discriminant validity. He argues that the high correlation among factors and the presence of a good second-order factor solution indicates that the objectives are not distinct. Similarity, our factor analysis of the common correlation matrix *across* rating forms suggests that in the data set as a whole, specific factors are not distinct. That is, different factors across rating forms often contain the same instructional categories. Thus, there is considerable overlap, and it does not make conceptual sense to treat the factors as distinct. If Marsh's arguments can be made for the discriminant validity of IDEA, surely they can also be made for the discriminant validity of the factors in student rating forms, including the SEEQ.

Thus, there are multiple indications, based on different rating forms, and utilizing different analytical techniques, that students rate instructors on the basis of three roles underling effective instruction. Three of the first-order factors of instructional effectiveness are highly correlated. This suggests that student ratings of instruction may contain a large *halo* effect. The traditional interpretation of such *halo*, is that it represents "errors" in judgement (Balzer, & Sulsky, 1992; Thorndike, 1920). However, more recent interpretations of *halo*, suggest not only that *halo* should not be considered error, but that it is a legitimate general factor (Lee, Malone, & Greco, 1981; Murphy, Jako, & Anhalt, 1993).

A second interpretation of *halo effect* in student ratings is that seemingly unrelated instructional characteristics are functionally and semantically related (Abrami, 1985; Cadwell, & Jenkins, 1985). That is, *dimensional similarity halo* would explain the factor invariance of student ratings. For example, effectively delivering instruction, may require the instructor to be enthusiastic, manage time well, cope with classroom disruptions, be sensitive to students' comprehension, and provide students' with meta-cognitive cues (feedback) to aid their comprehension. Thus, student rating of these behaviours would be consistently correlated since they are all part of the *meaning* of a good presentation. The results of this investigation, indicating that specific instructional behaviours load on three factors, *delivering instruction, facilitating interactions*, and *evaluating learning*,. also supports the view that *dimensional similarity halo* plays an important role when students rate their instructors.

Researchers (Feldman, 1988; Kishor, 1995; Whitely, & Doyle, 1976) investigating the *halo effect* in student ratings of instruction, have also suggested that *halo* may reflect the process of impression formation itself (Anderson, & Jacobson, 1965). Cognitive theories of *general impression halo* maintain that raters use common prototypes, emphasizing a person's role, to organize their impressions and subsequently recall or reconstruct performance ratings. Thus, according to these theories, students would organize their impressions of instruction on the basis of widely shared beliefs about what instructors do in general, and these general impressions would be subsequently activated by the items on student rating forms. The results of this investigation, indicating that *general instructional skill,* is a composite of delivering instruction, facilitating interactions, and evaluating learning, supports this view.

Since these beliefs would also be shared by faculty, rating constructors, and administrators, it is not surprising that there is agreement among student, peer, self, and supervisor ratings. Inter-rater discrepancies, on the other hand, may reflect the fact that students, faculty, and administrators hold different positions in the institutional hierarchy, and therefore, have both different knowledge of instruction and different goals. The reported factor invariance may also reflect selective hiring practices. Hiring committees may hire new faculty that share their perceptions of teaching.

Research has also suggested that *halo*, whether *dimensional similarity halo* or *general impression halo*, increases the accuracy of performance ratings (Murphy, Jako, & Anhalt, 1993). That is, global impressions may act as organizing principles, memory cues, *etc*., and improve students ratings of specific instructional behaviours. This suggests that global ratings (or a single score representing a weighted average of the specific ratings) rather than specific ratings should be used for personnel decisions.

Use of Multidimensional Rating Scores for Summative Decisions

While most researchers agree that teaching is multidimensional, they disagree on whether global or specific ratings should be used for summative decisions on retention, tenure, promotion, or salary. Abrami and d'Apollonia (1990) argued that since different student rating forms assess different dimensions of effective instruction, and since specific ratings are less reliable, valid, and generalizable than global ratings, they should not be used for personnel decisions. Rather, they suggested that a single score representing the average of several global items, or a carefully weighted average of several specific ratings be used. Marsh (1991, 1994) agreed with Abrami that a weighted averagc of specific dimensions is a good compromise. However, there is no consensus on how to weight the specific dimensions. The results of the present analysis speak to this issue.

The factor analysis across multiple forms indicates that specific dimensions are **not** distinct. Rather, they are chunked (by students while rating instructors) into three factors representing three instructional roles. Moreover, these factors are highly correlated and can be represented by a single hierarchical general instructional skill factor. As discussed above, students appear to base their ratings on both the semantic similarity among instructional dimensions (implicit theories) and on their overall impressions of the instructor.

Nathan and Tippins (1990, p. 291) argued that this "overall rating is not *halo* (error); it is a judgmental composite of what the rater believes is all of the relevant information necessary for making accurate ratings". Moreover, they reported that in a field study of performance ratings of clerical workers, the inclusion of a global rating increased the validity of specific ratings, defined as the correlation between supervisors' ratings and the clerical workers' scores on valid performance tests. They also reported that specific ratings added very little to the explained variance beyond that provided by the global rating. They concluded that, performance ratings are more efficient and accurate when based on global as opposed to specific ratings.

Similar results were reported by Cashin and Downey (1992) for student ratings of instruction. Here too, specific ratings added little to the explained variance beyond that provided by global ratings. Hativa and Raviv (1993) also reported that the global rating predicts from 90% to 61% of the variance in specific factor ratings. The factor analyses across the seventeen rating forms, reported in this thesis, indicate that whether one conducts a principal components analysis, an oblique rotation of the factor solution, or a second-order factor analysis, student ratings of instruction reflect general overall impressions. Thus, summative evaluations of instructors may also be more efficient and accurate when based on global rather than on specific instructional dimensions. Thus, in the next sections, I will investigate the validity of student ratings of instruction, viewed as one general hierarchical factor, general instructional skill, and explore the influence of study features on its validity.

PART III

THE VALIDITY OF STUDENT RATINGS OF INSTRUCTION

Literature Review

Although student ratings are used extensively in most post-secondary institutions, instructors and some researchers continue to express concerns especially when student evaluation is used for summative purposes. This is not surprising since researchers have reported validity coefficients for student ratings ranging from -0.93 (Endo, & Della-Piana, 1976) to 0.96 (Centra, 1977). The large variability in reported validity coefficients suggests that there are factors which modify the validity of student ratings. Unless these factors are known and accounted for, instructors and researchers will question summative decisions made on the basis of student ratings of instruction.

Definitions of Validity

The Standards for Educational and Psychological and Psychological Testing (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1985, p. 9) defines validity as referring to the "appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores". Although researchers usually refer to the validity of rating scales, tests, etc., it is the inferences made from these instruments that are validated, not the instruments themselves. Therefore, the validity of student ratings refers to the degree to which evidence supports the inferences on instructional effectiveness made as a result of student ratings.

Traditionally, the types of evidence gathered in support of these inferences have been labelled *content-related*, *criterion-related*, and *construct-related*. Content-related validity is a non-statistical estimate of the degree to which the content of an instrument corresponds to the content of the phenomenon it is designed to measure. That is, it demands that the items on the student rating form be a representative sample of items from the population of possible items. This also suggests that if a single rating form is to be used for summative purposes, it should contain items equally relevant to the classrooms in which it will be used. However, items tapping different dimensions pose a problem since not all dimensions are salient in all classrooms. For example, items assessing a warm, caring attitude may be more salient to Nursing students having difficulty with Physics, than to Engineering students wishing to be challenged. Wilson (1987) has also challenged the content validity of student ratings by arguing that student ratings evaluate instruction as defined by the text of the questionnaire. He argues that student ratings forms may reflect a particular view of education (top-down, authoritarian). Therefore, they may have content validity, but only as measures of an unjust and limiting definition of education. These student rating forms may not be appropriate in classrooms in which the instructor uses a more student-centred pedagogy, for example cooperative learning or discovery methods (Abrami, d'Apollonia, & Rosenfield, 1996).

Criterion validity is the degree to which "scores are systematically related to one or more outcome criteria" (American Educational Research Association, American Psychology Association, and National Council on Measurement in Education, 1985, p. 11). Criterion validity can be either concurrent or predictive; that is, the measurements on the instrument in question can be correlated to measurements on a criterion instrument for the same subjects at the same time or for the same subjects in the future. The estimation of criterion validity is relatively simple (the correlation of the instrument of interest with a criterion instrument). In general, the criterion validity of student ratings of instruction is determined by comparing student ratings with peer ratings, trained observer ratings, self ratings, or chair ratings, *etc*. Marsh (1987) reported, that in one of the few studies that correlated student ratings with instructor self-evaluation on the same instrument (SEEQ), the median correlation was .45 in one study and .49 in another study. Similarly, there is a high correlation between trained observers' ratings and student ratings (Murray, 1983; Marsh, 1987). On the other hand, peer or supervisor ratings (based on actual classroom visits) do not correlate highly with student ratings (Marsh, 1987).

The criterion validity approach to validating instruments has, however, been criticized (Leventhal, 1975; Messick, 1989; Zeller, 1988) on several grounds.

- Evaluations of criterion validity are restricted to the criterion used to establish validity (Zeller, 1988). For example, the criterion validity of student ratings established on the basis of trained observers does not necessarily indicate that student ratings will be valid measures for other criteria, such as student learning.
- The validity of the criterion measure against which the instrument of interest is being calibrated is itself questionable. Any irrelevant variance which contaminates the criterion instrument and is correlated to observations with the

instrument in question will cripple criterion validity approaches (Messick, 1989). For example, student ratings are usually based on more than 20 raters, while peer ratings may be based on only a few raters. Therefore, the reliability of peer ratings is much lower than that of student ratings, and attenuates the validity of peer ratings relative to student ratings (Marsh, 1987).

Construct validity is the degree to which the scores and their interpretation correspond to other measures of the same underlying theoretical trait (Cronbach, & Meehl, 1955). It subsumes both content and criterion validities (American Educational Research Association, American Psychology Association, & National Council on Measurement in Education, 1985: Messick, 1989). Establishing construct validity requires the following steps (Zeller, 1988):

- the explicit definition of the underlying theory, specifying the relationship of salient variables to the underlying trait (specifying a nomological network),
- selecting measures for the salient variables,
- describing the dimensional structure of the salient variables and scaling them appropriately,
- comparing the empirical correlations with the theoretically predicted correlations among the variables.

Although criterion validity is a necessary condition for construct validity it is <u>not</u> a sufficient condition. Critics of student ratings are concerned with questions of the construct validity of student ratings, not merely the criterion validity. The problem is that there may not be consensus on what teacher behaviours. in all cases. are causally related to effective instruction. In other words. there may not be an adequate theoretical model of effective instruction against which to evaluate the validity of student ratings. Zeller (1988) considers that the use of construct-related evidence to support the validity of an instrument requires that the nomological network surrounding the concept (s) be known. Thus, he considers construct validity approaches that rely heavily on factor analyses (*i.e.*, multitrait-multimethod approaches) to be inappropriate, since the intercorrelations can be contaminated by response sets.

In conclusion, the validity of student ratings of instruction should be based on collecting evidence from different sources to demonstrate that the inferences made as a result of the ratings are correct. When construct-related evidence is used, it is important to specify a model which includes both how instructors influence student learning and how students rate instruction. That is, one must demonstrate that the ratings are more closely related to instructional effectiveness (*i.e.*, student learning), than to other constructs such as student impressions, popularity. *etc.* .

Validation Designs

Four common methods of evaluating the validity of student ratings are multi-trait multi-method studies. true experimental studies. absence of *biasing factors* studies, and multisection validity studies. Since I will be integrating the multisection validity studies, I will briefly describe the first three designs, and describe the validation design in more depth.

Validation Designs Based on Multitrait-Multimethod Studies

The multitrait-multimethod approach to construct validation was developed by Campbell and Fiske (1959) who suggested that convergent and divergent validities could be assessed by measuring several different traits across several different methods. For example, courtesy, honesty, poise, and school drive can be measured by an association test or by peer ratings. The interpretation of test scores is valid if the correlations between measures that assess the same trait are higher (convergent validity) than the correlations between measures assessing different traits (divergent validities). Method or halo effects are inferred if the correlations among the different traits are higher when measured by the same method than when measured by different methods. However, for these inferences to be correct, all the measures must be equally reliable and valid. Although the MTMM method was initially designed to use maximally different traits measured by different methods, organizational psychologists began using it to measure different performance criteria as assessed by different raters (Borman, 1974).

The MTMM approach has been criticized on a number of grounds (Marsh, 1988). A very large number of comparisons must be made with no guidelines on how many comparisons must meet the Campbell-Fiske criteria to constitute evidence of construct validity. Halo or method effects contribute to convergent validity, and detract from divergent validity. However, there is no independent method of assessing whether halo is illusory or true. Moreover, it is highly unlikely that traits will be independent and measures equally reliable. Therefore, in the presence of halo or method effects, the interpretations of convergent and divergent validities are also ambiguous.

Kavanagh, Mackinney, and Wollins (1981) developed a three-factor (subject, trait, method) ANOVA model for analyzing MTMM data. When measures for all traits using all measures are obtained for the same subjects, three independent sources of variance can be obtained. The F-test on the main effects of subjects is used to infer convergent validity, the F-test of the subject x trait interaction is used to infer divergent validity, and the F-test on the subject x method interaction is used to infer halo or method effects. This approach has the advantage that it provides convenient summary statistics to infer construct validity. However, it too has the same limitations concerning the assumptions of uncorrelated traits and equally reliable measures as does the Campbell-Fiske approach. Moreover, the convergent, divergent, and method effects in the ANOVA model are not identical to those inferred from the Campbell-Fiske method (Marsh, 1988).

Marsh (1983a, 1983b, 1984) used the MTMM approach (both Campbell-Fiske and ANOVA models) to validate the SEEQ. He collected student and instructor self-evaluations and analyzed the scores across the nine specific factors. The nine convergent validities were on average .40, while the divergent validities were on average 0. However, there was also evidence of a large halo or method effect. For example, for student ratings of instruction, nearly 50% of the comparisons indicated the presence of a halo effect (Marsh, 1984). The ANOVA analysis led to similar conclusions. Both convergent and divergent validity coefficients were significant: however almost 20% of the variance component was attributable to halo effect.

This and similar MTMM approaches to the validation of student ratings (Howard, Conway, & Maxwell, 1985) have been criticized on a number of grounds (Gaski, 1987;

Abrami, d'Apollonia, & Rosenfield, 1996). For example, multiple criteria as opposed to a single criterion may amplify rather than reduce the problem of unreliable and invalid criterion measures. The construct validity approach requires that the different methods of assessing validity be "maximally different methods" (Campbell, & Fiske, 1959, p. 81); not different raters using the same instrument. Gaski (1987) argues such comparisons are better viewed as reliability estimates than as validity estimates. High convergent validities may equally support the conclusion that the alternative methods measure some other common latent trait (*eg.*, charisma) or a large halo effect than that they support the conclusion that the methods measure teaching effectiveness. Thus, Abrami, d'Apollonia, and Rosenfield (1996) concluded that the MTMM approach suffers from serious weaknesses and cannot provide unequivocal evidence for the validity of student ratings of instructional effectiveness. Unless, the multitrait-multimethod approach includes nomological validation (the degree to which predictions based on an explicit theoretical model conform to the observed correlations), this approach is mere "dust bowl empiricism" (Zeller, 1988 p. 329).

Validation Designs Based on True Experimental Studies

The original "Dr. Fox" study (Naftulin, Ware, & Donnelly, 1973) employed an actor who gave an "expressive" low-content lecture to students, who then favourably rated the instructor. Many researchers (Abrami, Leventhal, & Perry, 1982; Frey, 1979; Ware, & Williams, 1975) raised methodological and interpretive difficulties with the original "Dr. Fox" study. They subsequently developed a protocol for such laboratory designs in which an actor, exhibiting different levels of the instructor characteristic of interest (*eg,* high and low expressivity), presents a short video-taped lecture to students randomly assigned to treatments who subsequently both rate the instructor and take an achievement test. Thus, the treatment is the level of the instructor characteristic of interest. Meta-analyses of such laboratory studies (Abrami, Leventhal, & Perry, 1982) indicate that the influence of the instructor's expressivity on student learning was small ($\varpi^2 = .043$), whereas the influence on ratings was large ($\varpi^2 = .293$).

Abrami and his colleagues (Abrami, Leventhal, & Perry, 1982; Abrami, d'Apollonia, & Cohen, 1990; Abrami, d'Apollonia, & Rosenfield, 1996) have pointed out some of the shortcomings of laboratory studies as validation designs. For example, the range of the instructor characteristic of interest in laboratory studies does not represent the range found in actual instructors and since fixed levels of expressivity are selected, the laboratory results can not be extrapolated to field conditions. In addition, instructors affect both student learning and student ratings via a composite of many instructor characteristics, not only one. Manipulating only one or two instructional characteristics makes the characteristic artifactually salient. Thus, they concluded that laboratory studies are better used to demonstrate which teacher characteristics influence learning and/or student ratings, and to investigating the underlying causal mechanism. They are ineffective in determining the *degree* to which student ratings measure student learning.

Validation Designs Based on the Absence of Biasing Factors

In studies of factors which potentially *bias* student ratings, researchers infer that ratings are invalid (interpretations are contaminated with the influence of irrelevant factors)to the extent that student rating are correlated to these *irrelevant* factors and not to student learning, or *vice versa.* For example, if student ratings are moderately correlated to student sex, student ratings are interpreted as being invalid. However, as discussed by Abrami and Mizener (1985), Feldman (in press), and Marsh (1987), this interpretation is correct only if student learning is not causally affected by the *biasing* factor. If an instructor uses instructional techniques more attractive to males than to females, but these techniques have no effect on learning, ratings would be correlated to gender but not to achievement and therefore, gender would be a *biasing* factor. On the other hand, if gender were correlated to both ratings and learning, student rating forms would correctly reflect instructional effectiveness, and therefore gender would not be a biasing factor. In addition, the term *biasing* refers to those characteristics that change the correlation between student ratings and student achievement, and not to characteristics that render student ratings *unfair* (Feldman, in press). For example, untenured faculty may be given sections to teach which consist of unmotivated students who subsequently give their instructors low evaluations and do not learn much. Although it may not be appropriate to compare student evaluations given to untenured teachers with those given to teachers of classes consisting of motivated students, the ratings are valid. This validation design is discussed in greater length because *biasing variables* , by definition, moderate validity. Consequently, many of these variables will be coded as study features in the meta-analysis reported in this thesis.

There have been many studies on the influence of biasing factors, such as student, instructor, course and administration variables on student ratings. These have been extensively reviewed by Aleamoni (1981), Centra (1979), Feldman (1976, 1977, 1978, 1996), Kulik and McKeachie (1975), and Marsh (1987). The consensus has been that biasing variables play a minor role in student ratings of instruction. Marsh estimated that biasing variables account for between 12 to 14% of the variance in student ratings (cf. Murray, 1984).

Influence of student variables. Student sex, prior interest, expected grade, major, and level have been investigated as possible biasing factors by Feldman (1976, 1977, 1978). Of these, only student's prior interest, expected grade, and level have been consistently shown to affect student ratings. Feldman (1977), Haladyna and Hess (1993), Howard and Maxwell (1980), and Marsh (1987) reviewed previous studies investigating the influence of prior-interest on student ratings. Marsh (Marsh, & Cooper, 1981) found that prior interest was positively correlated to student ratings, but also to instructors' self-evaluation. Thus, they concluded that the correlation between prior interest and student ratings reflects a valid effect on student ratings and not a biasing factor. However, Marsh (1987) suggests that prior interest may reflect properties of the subject matter rather than of the instructor; and thus may have to be controlled across different courses if student ratings are to be used for summative decisions.

Many studies have found a moderate positive correlation between expected grades and student ratings (see Marsh, 1987). However, this finding can be interpreted in three radically different ways. It can be interpreted as providing evidence that instructors who give high grades get high evaluations (grading leniency hypothesis); that students who learn well from their instructors (and get high grades) evaluate their instructors appropriately (validity hypothesis); or that students who are motivated have prior interest or abilities, learn better, get better grades and evaluate their instructors appropriately (student characteristics hypothesis). Studies employing path analysis (Howard, & Maxwell, 1982, Marsh 1983) support the student characteristics hypotheses. Prior interest precedes expected grade and the covariance between expected grades and student ratings is eliminated by controlling for student motivation and other such student characteristics.

Abrami, Dickens. Perry, and Leventhal (1980) carried out a series of laboratory studies in which they experimentally manipulated grade expectations. They were able to show only a weak and inconsistent influence of grading standards on student ratings. Snyder and Clair (1979), in another laboratory study, demonstrated that the violation of grade expectations, rather than expected grades influenced student ratings.

A number of studies (*cf.* Feldman. 1978; Marsh 1980; Miller, 1972) have shown that students in advanced courses give more favourable ratings than do students in lower level courses. However. this effect is eliminated when the covariance between student level and student ratings is controlled by way of other student characteristics.

Influence of instructor variables. Instructor rank, experience. sex, research productivity and personality have been investigated as possible *biasing* variables. Of these. only instructor rank, research productivity, and personality have been consistently shown to affect student ratings. There have been a number of reviews that have examined the influence of instructor rank on student ratings (Centra, & Creech, 1976; Feldman, 1983; Marsh, 1987). Rank was significantly correlated with student ratings for specific teaching dimensions, but not for global ratings. That is, tenured faculty received higher scores than did teaching assistants on such dimensions as *Breadth of Coverage*, *Instructor Knowledge*, *Instructor Expansiveness* and lower scores on such dimensions as *Group Interaction*, *Encouragement of Discussion*, *Openness*, and *Concern for Students*.

Feldman (1988) carried out an extensive quantitative review of the literature on the influence of research productivity on student ratings of instruction. He found a positive weak correlation between research productivity and global ratings (.12) and higher positive correlations between research productivity and some specific dimensions, *Knowledge of the Subject Matter* (.21), *Preparation and Organization of the Course* (.19), *Clarity of Course Objectives and Requirements* (.18), and *Intellectual Expansiveness* (.15).

Feldman (1986) also extensively reviewed the association between 14 instructor personality characteristics and student ratings. He found that the correlations were both significant and large when the personality characteristics were inferred from student and colleagues ratings of instructor personality but insignificant and small when inferred from self reports. When instructor personality traits were *measured* by students or peers, most clusters of personality traits were significantly correlated to global ratings of teaching effectiveness. For example, the following personality characteristics (in descending order) explained more than 25% of the variability in global student ratings of instruction:

energy and enthusiasm; *positive view of others, i.e., tolerant, sympathetic, supportive, and warm*; *ascendancy, forcefulness, conspicuousness and leadership*, *reflectiveness, intellectuality, cultural and aesthetic sensitivity, flexibility, adaptability, openness to change, and adventurous*, *emotional stability, self-regard and self esteem*. Marsh (1987) pointed out that although Feldman's review indicated that instructor personality characteristics and instructional effectiveness are correlated; it did not indicate whether these personality characteristics were biasing factors. Moreover, research needs to be carried out on the influence of these personality characteristics on specific instructional dimensions.

Murray, Rushton, and Paunonen (1990) investigated the influence of instructor personality characteristics and student ratings across different types of psychology courses taught by the same instructor . They found that not only were peer ratings of personality traits good predictors of student ratings of instruction, but the pattern of instructor personality traits correlated to global student ratings of instruction differed significantly across course types. That is, the personality traits associated with effective teaching (as perceived by students) depended on the course type. For example,

- In large, lower-level, introductory courses, the effective instructor is likely to be "friendly, warm, and approachable, has a flair for the dramatic, and is fair and reasonable in relations with students, but shows an element of neurotic worrying" (Murray, *et al.*, 1990, p. 258).
- In smaller, higher-level discussion-oriented courses, the effective instructor is likely to be "friendly, gregarious, fair and supportive, and, at the same time,

flexible, adaptable, and open to change" (Murray. *et al.*. 1990. p. 258).

- In required honours and graduate courses. the effective instructor is likely to be "ambitious, competent, and hard working, and, at the same time, confident and worry free"(Murray, *et al.*, 1990, p. 258).

Murray *et al.* (1990) argue that the influence of personality traits is mediated through specific classroom behaviours, and that these behaviours affect student learning. Since specific rating dimensions reflect specific classroom behaviours. the implication is that the saliency of the specific rating dimensions would also vary across course types and therefore. although these characteristics do not *bias* student ratings, administrators need to take into consideration how they allocate courses to faculty.

Influence of course variables. Course status (elective or compulsory), class size, workload/difficulty. and academic discipline have been investigated as possible *biasing* factors. Of these, only course status, class size. and academic discipline have been consistently shown to affect student ratings. Research has generally shown that elective courses are rated more favourably than compulsory courses (Centra. & Creech. 1976; Feldman. 1978; Mintzes, 1977). However. Marsh (1987) argues that if prior interest is controlled for. this relationship is eliminated.

The influence of class size on student ratings has been extensively reviewed (Centra, & Creech, 1976; Feldman. 1978. Marsh, 1980). In general, there is a very weak negative correlation between global ratings and class size. However, class size was moderately negatively correlated with instructional dimensions pertaining to student-instructor interactions and rapport. However, if very large classes are included in the analysis, most researchers report a curvilinear relationship, with small and large classes evaluating instructors more favourably. Smith and Glass (1980) conducted an extensive meta-analysis of the research on class size and concluded that at class sizes between 1 and 40, student achievement, is negatively correlated with class size. This would suggest that student ratings reflect the true influence of class size on learning and that therefore class size is not a *biasing* factor. Moreover, Marsh (1987) argues that class size is correlated only with those dimensions that one would predict to be affected, and this supports the validity of these specific dimensions.

A number of researchers (Centra, & Creech, 1976; Neumann, & Neumann, 1985) and reviewers (Feldman, 1978) found that student evaluations were higher in soft disciplines such as the humanities than in hard disciplines such as mathematics and the physical sciences. Neumann and Neumann argued that these differences reflect the different roles of teaching in soft and hard disciplines. If this is true, it strongly suggests that specific instructional dimensions also play different roles in different academic disciplines and therefore should not be used for summative decisions when faculty are compared across disciplines.

Influence of administration variables. Anonymity, purpose of rating, and timing have been investigated as possible *biasing* factors. Of these, only timing has been consistently shown to affect student ratings (Feldman, 1978). Marsh and Overall (1980) reported that student ratings collected midway through the semester were substantially lower than ratings collected at the end of semester. Cohen (1981). on the basis of a meta-analysis of the multisection validity literature. reported that timing was a significant predictor of the size of the validity coefficient. Ratings collected at the end of the semester resulted in much higher validity coefficients than did ratings collected before students knew their final grade. This suggests that timing of evaluation will be a biasing factor.

In conclusion, some of the characteristics shown to be related to student ratings are prior interest, expected grades, reason for taking course, workload. level of course, class size, academic discipline, instructor rank and personality, and timing of evaluation. Marsh (1987) argues that there are a number of logical and methodological problems in interpreting variables as *biasing* variables, especially when they rely on correlational analysis in the absence of theoretical models specifying the relationships. However, this literature is discussed at length because the variables in question are potential moderators of the validity of student ratings.

Validation Designs based on Multisection Validity Studies

In the multisection validity design, first used by Remmers. Martin. and Elliot (1949), researchers correlate the section average score on student ratings with the section average score on a common achievement test across multiple sections of a multisection course. All sections use a common text book and common syllabus. Ideally, students are either randomly assigned to sections or ability is statistically controlled. This validity design has been criticized by Marsh (1984,1987) on a number of points (*eg.*, small sample size, poor randomization or statistical control of presage variables, small effect due to instructors when so many course characteristics are held constant, validity of the single criterion measure, and the possibility that students are rewarding lenient instructors with high ratings. Abrami, d'Apollonia and Cohen (1990) have responded to some of these criticisms by arguing that the multisection validity design has high internal validity and reduces threats to external validity (setting effects). For example,

- The use of the section rather than the individual average scores emphasizes the role of the instructor.
- The use of the common final exam as the criterion variable increases reliability by reducing the influence of differential instructor grading practices.
- The common achievement test is also likely to have appropriate validity in the types of courses used in multisection courses (*i.e.*, large introductory courses).
- Random assignment or statistical control reduces the chances that extraneous student characteristics will act as *biasing factors*.
- The attempt to reduce section to section differences by standardizing course features such as syllabus, objectives, and textbook also reduces the probability that course, rather than instructor characteristics, will bias student ratings.

Leventhal (1975) suggested that the criterion validity approach of the multisection validity design cannot demonstrate the construct validity of student ratings because it relies on predictive relationships rather than on causal relationships between student ratings and learning. He suggested that the criterion validity approach could be strengthened by reducing the problem of section inequivalences (by increasing experimental and/or statistical control), by demonstrating that student ratings are correlated with intermediary variables which are causally related to instructor characteristics which affect learning, and by the use of panel designs. Abrami, d'Apollonia, and Rosenfield (1996) suggest that the multisection validity design can be improved by including multiple measures of student learning and by empirically examining the links between instructional behaviours and student learning.

Researchers have argued about the relative merits of validation designs (Abrami, d'Apollonia, & Cohen, 1990; Gaski, 1987; Howard, Conway, & Maxwell, 1985; Marsh, 1987; Maxwell, & Howard, 1987). However, they often argue on the basis of different criteria for validity. Comparing the methodological quality of different validation designs is analogous to comparing the methodological quality of primary studies in a meta-analysis. Although methodological quality of primary studies can be assessed easily on the basis of an individual criterion, it is extremely difficult if not impossible to accurately and objectively make summative evaluations on the basis of the set of criteria (L'Hommedieu, Menges, & Brinko, 1987). Similarly, methodological quality of validation designs is a summative evaluation of the "degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inference and actions based on test scores" (Messick, 1989 p. 5). Arguments on the rival merits of a validation design high in internal validity but low in construct validity, for example, compared to a validation design weak in internal validity but strong in construct validity are inconclusive without prior agreement on the importance of the respective criteria.

In conclusion, although the multisection validity design is not perfect, it is a validation design that has high internal validity. Moreover, many of its flaws are not inherent to the design, but rather decisions that researchers take. For example, statistical control of presage variables can be included, larger sample sizes can be selected, *etc.*. To date, more than fifty studies have utilized the multisection validity design to determine the validity of student ratings. In addition, there have been numerous reviews, both qualitative and quantitative. Because this validation design has been used so extensively, with many student rating forms under diverse conditions, it provides the most generalizable evidence for the validity of student ratings.

Validity of Student Ratings as Determined by Multisection-Validity Studies

The multisection validity literature has been extensively reviewed both quantitatively (Abrami, 1984; Cohen, 1981, 1982, 1983; d'Apollonia, & Abrami, 1988; Dowell, & Neal, 1982; Feldman, 1989, 1990; McCallum, 1984) and qualitatively (Feldman, 1978; Kulik, & McKeachie, 1975; Marsh, 1987). However, not only do the primary researchers reach opposite conclusions; so do the reviewers. For example, Cohen (1981) concluded that student ratings were valid measures of teaching effectiveness; while Dowell and Neal (1982) and McCallum (1984) concluded that ratings, at best, were poor measures of teaching effectiveness. The Cohen (1980, 1981) meta-analysis was the most complete review of the multisection validity literature and has formed the basis of subsequent meta-analyses (Feldman, 1989; Abrami, d'Apollonia, & Cohen, 1990). Therefore, the results of this meta-analysis are described in greater detail than are the other reviews.

Cohen Meta-analysis

Cohen (1980, 1981) extracted and coded 266 validity coefficients from 41 studies into ten instructional categories: two global dimensions assessing the overall course and instructor effectiveness; six specific instructional dimensions based on Kulik and McKeachie's (1975) "common factors" (*Skill, Rapport, Structure*, and *Difficulty*); Issacson *et al's*. (1964) factor study (*Interaction* and *Feedback*); and two additional specific instructional dimensions (*Evaluation* and *Learning Progress*). He appears to have extracted only one validity coefficient per class section per category, although no decision rules are specified. Table 13 presents the ten unweighted mean effect magnitudes computed by Cohen (1981) using the procedures described by Glass (1978). The three global ratings and two specific ratings (*Skill* and *Structure*) had moderately high validity

Table 13. *The mean validity coefficients of student ratings (from Cohen, 1980).*

Instructional Category	N	Mean validity
Overall Learning Progress	14	.47
Overall Course	22	.47
Overall Instructor	67	.43
Skill	40	.50
Structure	27	.47
Rapport	28	.31
Feedback	5	.31
Evaluation	25	.23
Interaction	14	.22
Difficulty	24	-.02

Note N is the number of validity coefficients.

coefficients. Four dimensions. *Rapport*. *Feedback*. *Evaluation*. and *Interaction* had moderate to low validity coefficients (.31 to .22); while the validity coefficient for *Difficulty* was 0.

Cohen (1980, 1981) was the only reviewer to attempt to explain the variability in reported validity coefficients. He analyzed the following three sets of study features:

- methodological features (student assignment, statistical control of student ability type of achievement test, source of achievement test, evaluation of student achievement, instructor prior knowledge of achievement test, source of rating instrument, timing of student evaluation, length of instruction, instructor autonomy, number of sections, overall study quality);
- ecological conditions (institutional setting, instructor experience, course level);
- 3 course characteristics (i.e., pure or applied, hard or soft, life or non-life); and
- publication features (source, publication year).

He analyzed the influence of these twenty study features on the global *Overall Instructor* category and on the four specific instructional categories (*Skill* .*Rapport*. *Structure*. *Difficulty*). The results of this analysis are presented in Table 14. Because of differences in sample sizes and therefore differences in statistical power (Abrami. Cohen. & d'Apollonia. 1988), study features having large to medium effects on the validity coefficients do not necessarily reach significance. For example. type of achievement test (objcctive *vs* not objective) is a significant predictor for the *Overall*

Table 14. *The correlations greater than ± .20 between study features and the mean validity coefficients (from Cohen, 1980).*

Study Feature	Overall Instructor	Skill	Rapport	Structure	Difficulty
	N=67	N=40	N=28	N=27	N=24
timing of evaluation	-.43 *	-.56 *	-.02	-.27	-.45 *
scoring of final test	.12	.06	.03	.31	.01
source of final test	.12	-.12	-.02	-.27	.04
type of final test	.29 *	.35 *	.29	.37	.08
number of sections	-.14	-.16	-.05	-.64	-.02
overall quality	-.04	-.11 .	-.00	-.42 *	-.13
instructor experience	.25 *	.25	.20	.23	-.12
hard discipline	-.01	.21	.26	-.04	.09
life discipline	-.06	-.23	-.11	-.03	-.08
course level	.13	.20	.08	.22	.33
institutional type	-.06	.00	-.20	.01	.02
publication year	.10	.26	.03	.08	-.06

Note N is the number of validity coefficients.

Instructor Rating and *Skill* rating, but not for *Rapport* and *Structure*.

Cohen, then conducted a hierarchical multiple regression analysis, initially with seventeen of the study features entered as three sets, and identified the significant predictors (given the specific order of entry) of the *Overall Instructor* rating. These were *instructor experience*, *timing*, and *evaluation bias*. These three predictors accounted for 31% of the variance in the validity of the *Overall Instructor* rating. Thus, the mean validity coefficient was .34 when the instructors were graduate students compared to .48 when the instructors were full-time faculty; .85 when the evaluation was carried out after students knew their final grade compared to .38, when they did not; and .15 when instructors graded their own students compared to .52 when external graders were used.

However. a large proportion of the variance (69%) remained to be explained.

Reanalyses of Cohen Meta-analysis

Subsequent reviewers such as Abrami, Cohen, and d'Apollonia (1988) and Abrami. d'Apollonia. and Cohen (1990) analyzed the quantitative reviews to explain why reviewers came to opposite conclusions. They identified a number of decisions made by the reviewers that *biased* their conclusions. For example. the reviewers differed on their inclusion criterion, and therefore the studies they included in their analysis. on the number of outcomes they extracted, on the analytical techniques used. *etc.*. They concluded that the quantitative reviewers differed both technically and conceptually and made the following recommendations for future quantitative reviews of the multisection validity studies:

- Reviewers should improve data analysis by selecting an appropriate analytical approach (traditional, homogeneity. or variance partition). by avoiding Type II errors. and by taking interdependencies among outcomes into account.
- Reviewers should improve the coding of outcomes and explore interrelationships among these outcomes.
- Reviewers should improve the coding of study features.

Data analysis. d'Apollonia and Abrami (1987, 1988) compared three analytical methods of conducting a meta-analysis: the traditional (Glass, McGaw, & Smith, 1981), the variance partition (Hunter, Schmidt, & Jackson, 1982), and the homogeneity approaches (Hedges, & Olkin. 1985). They conducted the three types of meta-analyses of the Cohen data set, with corrections for inclusion errors and additional studies indicated in Abrami, Cohen, and d'Apollonia (1988), and found that the results and therefore the conclusions, varied with the method used. They demonstrated that the variance partition and homogeneity approaches were more powerful than the traditional approach. Since there were few differences between the variance partition and homogeneity approaches. they recommended that the more widely used homogeneity approach be used in future analyses.

d'Apollonia and Abrami also investigated the choice of the unit of analysis. Meta-analysts can choose one validity coefficient per outcome category per study (the study is the unit of analysis), one validity coefficient per outcome category per multisection course (the course is the unit of analysis). or all the validity coefficients (the finding is the unit of analysis). They found, not unexpectedly, that statistical power is much greater when the finding is the unit of analysis[14]. However. more importantly. choosing the finding as the unit of analysis allows for within-study comparisons. For example. it allows the direct comparison (within one study) of the validity of student ratings when the raters are female *versus* when they are male. However. it does introduce problems associated with non-independence of within-study validity coefficients. Thus. Abrami. Cohen. and d'Apollonia (1988) called for further quantitative reviews of the multisection validity literature which would address issues of data extraction. coding of study features,

[14] Note that there are approximately 500 more validity coefficients when the finding is the unit of analysis than when the course is the unit of analysis in the Cohen. (1980) meta-analysis.

and new data analysis procedures dealing with within-study comparisons (dependent data). They suggested that homogeneity tests (Hedges, & Olkin, 1985) and multivariate procedures (Raudenbush, Becker, & Kalaian, 1988) be used in future analyses.

Coding of outcomes. One limitation of the Cohen (1980, 1981) meta-analysis was how validity coefficients were classified. Cohen used only ten categories and it was not clear how decisions were made to extract outcomes and fit them into instructional categories. Over the past two decades, Feldman (1976, 1983, 1984, 1989) developed an extensive coding schema to characterize the items used in multidimensional rating forms. This modified coding schema has been used in several meta-analyses of the multisection validity studies (d'Apollonia, & Abrami, 1988; Feldman, 1989).

d'Apollonia and Abrami (1988) coded the outcomes extracted from 44 studies in the modified multisection validity set (Abrami, Cohen, & d'Apollonia, 1988; d'Apollonia, & Abrami, 1988). Many of the validity coefficients come from factor scores that include more than one instructional dimension (Abrami, & d'Apollonia, 1990), and therefore were included more than once. Feldman (1976, 1983, 1984) corrected for this inflation in total number of entries by weighting each outcome by the inverse of the number of entries. Since we were not convinced of the appropriateness of this weighting procedure[15], and since each instructional category is analyzed separately, *i.e.*,

[15] Feldman stated that he weighted these multiple entries by the inverse of the number of entries to prevent these outcomes unduly influencing " the results of averaging within instructional dimensions" (Feldman, 1989, p 588). This weighting gives less credence to validity coefficients that are " of mixed dimensional structure" relative to "pure" validity coefficients. Moreover, this weighting scheme weights the dimensions equally regardless of how many items it contains. For example, Feldman would weight a rating consisting

univariately, we decided not to weight the outcomes at this stage. The results of this univariate analysis, presented in Table 15, indicate that in general, global ratings of instruction have higher validities than do specific ratings. The validity of specific ratings varies from a high of .31 (for *Clarity*) to a low of 0 for *Intellectual Expansiveness.*

Feldman (1989, 1990) also conducted a meta-analysis of the multisection validity literature. He used different criteria for the extraction of validity coefficients. For example, he only included validity coefficients computed from multidimensional rating forms and excluded any validity coefficients for scores based on more than six instructional categories. Feldman also weighted multidimensional outcomes by the inverse of the number of entries prior to aggregation. The results of his univariate analysis are also presented in Table 15. Although the validity coefficients, are somewhat higher, the general pattern of mean validities is similar across both the d'Apollonia and Abrami (1988) and Feldman (1989) analyses. For example, Table 15 indicates that in both analyses, *Organization, Clarity*, and *Interest* are among the five specific ratings that have the highest validity; while, *Respect* and *Difficulty* are among the five specific ratings that have the lowest validity.

Differences in inclusion criteria may account for some of the differences in reported results. For example, Feldman included fewer validity coefficients in his analysis than did d'Apollonia and Abrami. When his data set is compared to the

of 8 items in one dimension (X), and one item each in dimensions (Y) and (Z) equally (1/3) in each of dimensions X, Y, and Z. Surely, the weights should reflect the number of items, and therefore the entries should be weighted unequally.

Table 15. *The mean validity coefficients of instructional effectiveness from d'Apollonia, and Abrami (1988) and Feldman (1989).*

Instructional Category (Bolded labels refer to Feldman's categories)	d'Apollonia. & Abrami (1988)		Feldman (1989)	
	N	validity	N	validity
Clarity/ **Clarity and understandableness**	66	.31	32	.56
Interest/**Teacher's stimulation of interest**	50	.31	19	.38
Availability/ **Teacher's availability**	37	.25	22	.36
Organization/**Preparation and Organization**	63	.21	28	.57
Personality/ **Personality**	6	.18	6	.24
Feedback/ **Instructor's Feedback**	30	.11	21	.23
Course content/ **Course Material**	35	.16	17	.17
Responsiveness/ **Concern with Class Level**	39	.16	21	.30
Inst. Objectives/ **Instructor met Objectives**	50	.13	8	.49
Challenge/ **Motivated to excellence**	39	.13	3	.38
Discussion/ **Encouragement of discussion**	55	.12	28	.36
Enthusiasm /**Enthusiasm**	21	.11	10	.27
Fairness of Feedback/ **Instructor's Fairness**	54	.18	26	.26
Supplementary Materials/ **Supplementary Materials**	25	.10	5	-.11
Knowledge/ **Teacher's Knowledge**	23	.06	10	.34
Elocution/ **Teacher's Elocutionary Skills**	13	.05	6	.35
Respect/ **Respect and Friendliness**	55	.05	12	.23
Difficulty/ Difficulty of Course: **Descriptive**	52	.03	21	.09
Management/ **Management**	41	.02	5	.26
Intellect/ **Intellectual Expansiveness**	5	0	2	.04

Note N is the number of validity coefficients.

d'Apollonia and Abrami data set, many of the excluded findings in the Feldman data set had low validity coefficients. Differences in weighting procedures would also have resulted in higher mean validity coefficients for the Feldman procedures. According to Benton and Scott (1976), validity coefficients for multidimensional rating scales are attenuated. Feldman (1989) would have weighted validity coefficients from multidimensional scales less than those from unidimensional scales; while, d'Apollonia and Abrami (1988) weighted them equally.

Therefore, there is general agreement that for most instructional dimensions, instructional effectiveness (as judged by student ratings) is positively correlated to student achievement. However, there is a great variability in the magnitude of the association. Such instructional dimensions as *Instructor Clarity, Interest*, and *Organization* are highly correlated to student learning; while *Course Difficulty, Management*, and *Respect and Friendliness* have little association with student learning. In addition, both d'Apollonia and Abrami (1988) and Feldman (1989) demonstrated that the method of averaging and the unit of analysis influence the results of the meta-analysis, with the methods used by Cohen giving the largest validity coefficientso.

Coding study features. Abrami, d'Apollonia, and Cohen (1990) nomologically coded the study features that were investigated, accounted for, and mentioned by primary researchers. They reported that of the 520 variables described by primary researchers, 35% were investigated, 48% were accounted for, and 18% were only discussed. In addition, they reported that 954 additional variables were mentioned in the primary studies. They then categorized the 520 variables into 75 categories grouped into four sets: rating characteristics, achievement characteristics, explanatory characteristics, and miscellaneous features. They reported that 35% of the 520 variables were rating characteristics, 12% were achievement characteristics, 31% were explanatory characteristics, and 23% were miscellaneous characteristics.

Abrami, d'Apollonia, and Cohen (1990) also nomologically coded the study features reported by the reviewers of this literature using the same four-category schema.

The reviewers reported 199 study features of which 16% were rating characteristics. 13% were achievement characteristics. 56% were explanatory characteristics. and 22% were miscellaneous characteristics. A comparison of the frequency distributions of the 75 categories reported by the primary investigators and the reviewers indicated that the frequencies were significantly different. In general. researchers often investigated characteristics of ratings but seldom investigated explanatory characteristics. On the other hand. the reviewers frequently reported explanatory characteristics. Furthermore. the researchers were primarily interested in student explanatory characteristics. while the reviewers emphasized the instructional explanatory characteristics. Thus. Abrami. d'Apollonia. and Cohen called for further primary studies investigating some of the poorly-researched characteristics. They also suggested that this nomological coding schema be used in future quantitative reviews of the multisection validity literature.

Implications of Literature to Research

The multisection validity studies indicate that global ratings of instruction are valid. accounting for 10% or more of the variance in student achievement. The validity of specific ratings of instruction is more diverse. with some specific ratings (*e.g.*, *Preparation. Clarity. Interest*) accounting for more than 10% of the variance in student achievement: and other specific ratings (*e.g.. Difficulty. Course Materials. Feedback*) having little or no association with student learning. However. the findings for *both* global and specific ratings are heterogeneous, suggesting that study features may moderate the strength of the association between student ratings and student achievement.

Some study features (*Timing, Instructor Rank*, and E*valuation Bias*) have been shown to moderate the validity of student ratings. However, much of the variability in findings remains to be explained. Thus, the goal of PART III of this thesis is to apply some of the newer more sophisticated meta-analytical techniques (weighting by sampling variance, modelling the interdependencies among outcomes), and to attempt to explain the factors affecting the validity of student ratings of instruction.

Methods

The five steps in integrating the validity studies are: problem formulation and specification of inclusion criteria, identification of studies, extraction and calculation of validity coefficients, coding outcomes and study features, and data analysis. These five steps provide the framework for the description of the methodology.

Problem Formulation and specification of inclusion criteria

I addressed three questions in this meta-analysis. The first question was *Do moderator variables differentially affect the validity of the four primary factors of instructional effectiveness ?* In other words, are there significant and practically important interactions between moderator variables and the factor structure of student ratings. If there are, I will investigate the validity of student ratings across the four primary factors; otherwise, I will investigate the validity of student ratings across the general skill factor identified by the secondary factor structure. The second question was *What is the overall validity of student ratings as measures of instructional effectiveness reported in the* *multisection validity studies?* The third question was *To what extent is the multisection validity literature consistent, and if it is not consistent, to what extent do study features explain the variability in reported validity coefficients ?*

The operational definition of instructional effectiveness in this study is the mean class performance on an achievement test. Validity is defined as the strength of the relationship (Pearson-product correlation coefficient) between class section mean student ratings of instruction and class section means on a common achievement test. The inclusion criteria for this study were the following:

- All studies must take place in post-secondary settings. Thus, military studies (*e.g.*, Borg, & Hamilton, 1956; Chase, & Keene, 1979) were excluded.
- All studies must come from actual classes and not simulated classes.
- All studies must report validity coefficients across multi-sections of the same course or provide data which can be used to compute validity coefficients.
- Validity coefficients must be based on section means of achievement and ratings.
- All studies must use a common criterion of student achievement across sections. Thus, Bendig (1953b) was excluded.

Identification of Studies

The primary sources of studies for the meta-analysis were the previously published quantitative reviews (Abrami, 1984; Cohen, 1981; Cohen, 1982, 1983; Dowell, & Neil, 1982; McCallum, 1984). The following steps were carried out:

- Computer searches were carried out of Psychological Abstracts, Educational

Resources Information Centre (ERIC), Comprehensive Dissertation Abstracts, Social Citation Index, Sociological Abstracts, Medlines, National Technical Information Services, and Inspec using the key words described by Cohen (1981). For the ERIC search these were: (course evaluation or student evaluation of teacher performance or teacher evaluation) and (academic achievement or grades scholastic) and (higher education or colleges or postsecondary education or universities or college instruction).

- Recent and unpublished articles were solicited through an announcement in *Instructional Evaluation* and by correspondence with seventy-two active researchers in the area.
- Manual searches were carried out on recent journal issues and annual programs of the American Educational Research Association and the International Conference in Improving University Teaching.
- The above steps were carried out up to December 1987 (Abrami, Cohen, & d'Apollonia, 1988). The computer searches of ERIC and Psychological Abstracts were updated to December 1996.

The above strategy uncovered four additional studies that met the inclusion criteria (McKeachie, Lin, & Mendelson, 1978; Morgan, & Vasche, 1978; Murray, 1983; and Soper, 1973). All the identified studies were acquired, with the exception of Spencer and Dick (1965), which was not available. Thus, there are 43 studies in the final data set.

Extraction and Calculation of Outcomes

Two people extracted 741 validity coefficients (Pearson product moment correlation, Spearman correlation, and Kendall correlation). from the 43 studies. The interrater reliability was .85 (Cohen's kappa). The major source of disagreement was the extraction of validity coefficients for both separate factors (or items) and for the sum (or average) across items or factors. All disagreements were resolved by mutual agreement.

The equation (Gibbons, 1971, p. 232) used to convert Spearman's *rho* to the standard normal deviate was the following:

$$Z = \rho\sqrt{n - 1}$$

where **Z** is the standard normal deviate,
ρ is Spearman's rho,
n is the number of sections .

The equation (Gibbons, 1971, p. 218) used to convert Kendall's *tau* to the standard normal deviate was the following:

$$Z = \frac{3\sqrt{n(n - 1)} \cdot \tau}{\sqrt{2(2n - 5)}}$$

where **Z** is the standard normal deviate,
τ is Kendall's *tau*,
n is the number of sections.

The equation (Rosenthall, 1994, p 239) used to convert the standard normal deviate to the Pearson product-moment correlation coefficient was the following:

$$r = \frac{Z}{\sqrt{n}}$$

where **Z** is the standard normal deviate,
r is the Pearson product-moment correlation coefficient,
n is the number of sections.

The equations (Shadish, & Haddock, 1994, p. 268) used to transform the Pearson product moment correlation coefficient to Fisher's z and back were the following:

$$z = .5 \cdot \ln\frac{(1 + r)}{(1 - r)}$$

$$r = \frac{(e^{2z} - 1)}{(e^{2z} + 1)}$$

where **z** is Fisher's z,
r is the Pearson product- moment correlation coefficient,
n is the number of sections.

Coding Outcomes and Study Features

In this meta-analysis, the outcomes are the validity coefficients reported by the primary researchers. The validity coefficient is the correlation between section average scores on student ratings and section average scores on a common achievement test across multiple sections of a course. Since the rating forms used in the different studies can vary widely, I coded the outcomes as well as the study features.

Coding Outcomes

I subsequently coded the extracted outcomes to reflect the items on the student rating forms that were used to compute the outcomes (validity coefficients). These outcomes were coded at three levels:

- to reflect the contribution of the individual categories;
- to reflect the contribution of the first-order factors; and,
- to reflect the contribution of the second-order factors.

In each case, they can be coded either continuously (*e.g.*, as the proportion of items within each category) or categorically (*e.g.*, as the presence or absence of items within each category). Note that some of these outcomes are coded as intermediate values in order to code other variables. The following four outcome variables were coded:

- (PITEM's): Thirty-five continuous variables ($PITEM_1$, $PITEM_2$, ..., $PITEM_{35}$) were computed. The items that contributed to each validity coefficient were listed and the proportion of items fitting into each of the thirty-five categories identified in the previous meta-analysis were recorded.
- (PFAC's): Four continuous variables ($PFAC_1$. $PFAC_2$. $PFAC_3$. and $PFAC_4$) representing the proportion of items contributing to each first-order factor were computed. These values were subsequently used as a study feature representing the "structure" of the rating forms. The (category by factor) factor structure (**S**) matrix was used to determine the factor on which each category loaded the highest. It is calculated from the following formula:

$$S = P\,\Phi$$

where **S** is the (category by factor) structure matrix,
P is the (category by factor) pattern matrix,
Φ is the (factor by factor) factor correlation matrix.

A (category by factor) weight matrix (**W**) was constructed by entering for each category, a 1 for the factor on which the category loaded the highest, and a 0 on all other factors (Gorsuch, 1983). The proportion of items contributing to each factor (**PFAC**) for each outcome was calculated by the following formula[16]:

$$F = W\,D$$

where **F** is the row of factor scores on the four factors,
W is the (category by factor) weight matrix,
D is a row of PFAC variables (the proportion of items in each category).

- (PHFAC's): Two continuous variables (PHFAC_1 and PHFAC_2) representing the proportion of items contributing to each second-order factor were computed. A (category by factor) weight matrix ($\mathbf{W}_{II}$) was computed using the following formula from Gorsuch (1983, p. 247):

$$W_{II} = PP_{II}$$

where $\mathbf{W}_{II}$ is th(category by second-order factor) weight matrix,
P is the (category by first-order factor) pattern matrix,
$\mathbf{P}_{II}$ is the (first-order by second-order factor) factor loading matrix..

[16] This can also be calculated using the following formula:

$$f_i = \sum_{i=1}^{k} w_i p_k$$

where fi is the proportion of items contributing to each factor score,
k is the individual category,
w is the weight assigned to each
p is the proportion of items in each category.

The proportion of items contributing to each second-order factor (PHFAC) for each validity coefficient was calculated by the following formula[17]:

$$F = W_{II}D$$

where **F** is the row of factor scores on the four factors,
W_{II} is the (category by second-order factor) weight matrix,
D is a row of PFAC variables (the proportion of items in each category).

- (HFAC's): Two categorical variables ($HFAC_1$ and $HFAC_2$) representing the presence or absence of the two second order-factors were calculated by giving a value of 1, if the HPFAC was greater or equal to .10, and 0. if the value was less than .10.

Coding Study Features

Since the unit of analysis, in this thesis, is the validity coefficient rather than the study, the characteristics of the validity coefficient, not the study. must be coded. Nomological coding was used to select and organize the study features in the multisection validity studies (Abrami, d'Apollonia, & Cohen, 1990). The coding schema used to code the study features in this paper. is presented in Appendix 3. Mean substitution was used

[17] PHFAC's can also be computed by the following formula:

$$f_i = \sum_{i=1}^{k} w_{H_i} p_k$$

where f_i is the proportion of items contributing to each second-order factor score,
k is the individual categories
w_{II} is the hierarchical weight assigned to each category,
p is the proportion of items in each category.

for missing data for continuous variables missing less than 10% of the values. Continuous variables that had more missing values were categorized by either a median or tertiary split. Missing values for the categorical variables were coded.

In this thesis, I explored the extent to which study features moderate the validity coefficient. That is, I am exploring whether information about study features explains the variability in reported validity coefficients. I therefore "dummy coded" the categorical variables into p-1 vectors (where p is the number of levels of the variable) with missing values coded as 0 in all vectors.

Data Analysis

The description of the data analysis is subdivided into three sections: the synthesis of the variance-covariance matrix which models the dependencies among outcomes, the calculation of the population parameters, and the general linear model used to test the influence of study features. All analyses were carried out in MATRIX PROC, within SPSS (SPSS. Inc., 1994).

Synthesis of the Variance-Covariance Matrix

Since there are multiple outcomes from many of the studies, the data is interdependent. This dependency was modelled using the methods described by Becker (1992), Gleser and Olkin (1994), and Raudenbush, Becker, and Kalaian (1988) . In these procedures, the off-diagonal elements of the variance-covariance matrix include non-zero elements between outcomes that are dependent and zero entries between independent outcomes. These off-diagonal elements are the correlation coefficients between dependent outcomes. calculated as described by Becker (1992. p. 235).

$$\sigma_{i,j} = \frac{r_{ic} \cdot (.5 \cdot (2 \cdot r_{ic}) - (z_i \cdot z_j)) \cdot (1 - (z_i \cdot z_i) - (z_j \cdot z_j) - r_{ic}^2)}{n_i}$$

where $\sigma_{i,j}$ is the covariance between two dependent validity coefficients from one study,

$\mathbf{r}_{ic}$ is the correlation coefficient between two dependent outcomes.

$\mathbf{z}_i$ is one dependent Fisher transformed validity coefficient.

$\mathbf{z}_j$ is the second Fisher transformed validity coefficient,

$\mathbf{n}_i$ is the number of sections used to calculate the validity coefficients.

Since, the outcomes in this study reflect subsets of items contributing to one of the second-order factors (Hierarchical Factor 1), we used the internal consistency of this factor as a measure of the correlation coefficient $\mathbf{r}_{ic}$.

The diagonal elements of the variance-covariance matrix are the sampling variances associated with each outcome. The formula for the sampling variance for an individual outcome, taken from Hedges and Olkin (1985, p. 231) is the following.

$$\sigma^2_{z_i} = \frac{1}{n_i - 3}$$

where σ^2_{zi} is the sampling variance for each Fisher- transformed validity coefficient.

$\mathbf{n}_i$ is the sample size (number of sections) associated with each outcome.

Since in the multivariate methods. the validity coefficients are weighted by the *inverse* variance-covariance matrix, the variance-covariance matrix for each independent set of data was inverted using SPSS PROC MATRIX (SPSS,1994).

Finally, the validity coefficients computed for multidimensional ratings can enter the analysis more than once, inflating the sample size. This, can be corrected by weighting the validity coefficients by the inverse of the number of replications (see Feldman, 1989). In this investigation, I used the proportion of items per rating form that contributed to the validity coefficient as the weighting factor. Since, some categories did not contribute to the factor analysis and since only one of the two second-order factors was analyzed, the sum of weights can lead to a reduction in the degrees of freedom, relative to the unweighted analysis.

In conclusion, a weighted inverted variance-covariance matrix was synthesized for each set of within-study dependent validity coefficients. Each submatrix was subsequently combined to form the weighted inverted variance covariance matrix for the complete data set. Note that all off diagonal elements between the submatrices will be 0 in this 741 by 741 matrix. If a different subset of the data is to be analyzed, a new weghted inverted variance covariance matrix must be synthesized of the appropriate rank.

Calculation of Population Parameters

The population parameters that were calculated and are reported here as the mean effect magnitude (mean validity coefficient), the standard error used to compute the 95% confidence limits, and the homogeneity statistic. All calculations were carried out using SPSS PROC MATRIX (SPSS, 1994).

Mean effect magnitude. The mean effect magnitude, weighted by the inverted variance-covariance matrix, was computed using the following formula described by

Becker and Schram (1994, p. 369) but using z rather than *r* as recommended by Hedges and Olkin (1985):

$$MEM_z = X^t\sum^{-1} X^t\sum^{-1} z$$

where **MEM$_z$** is the Fisher transformed mean effect magnitude or mean validity coefficient;
X^t is a design matrix representing the predictors in a multiple regression model. In this case, it is a row vector containing **n** 1's,
n is the number of validity coefficients that are being aggregated;
$\sum^{-1}$ is the inverted variance-covariance matrix, described above;
z is a column vector of the Fisher transformed validity coefficients.

Standard error and 95% confidence interval. The standard error was computed using the formula described by Becker and Schram (1994, p. 369):

$$SE_{MEM_z} = \sqrt{((X^t\sum^{-1} X)^{-1})}$$

where **SE$_{MEMz}$** is the standard error of the Fisher transformed mean effect magnitude or mean validity coefficient;
X^t is the predictor matrix representing the predictors in a multiple regression model. In this case, it is a row vector containing **n** 1's, where **n** is the number of independent validity coefficients that are being aggregated;
$\sum^{-1}$ is the inverted variance-covariance matrix, described above.

The 95% confidence interval is calculated using the following formula:

$$95\%CI = MEM_z \pm z_{\frac{\alpha}{2}} SE_{MEM_z}$$

where **MEM$_z$** is the Fisher transformed mean effect magnitude or mean validity coefficient;
SE$_{MEMz}$ is the standard error of the Fisher transformed mean effect magnitude or mean validity coefficient;
z$_{\alpha/2}$ is the 100th (1 - α/2) percentile of the standard normal distribution.

Homogeneity statistic. The homogeneity statistic was computed using the formulae described by Hedges and Olkin (1985 p. 211).

$$Q_T = z' M z$$

where **z** is a column vector of the Fisher transformed validity coefficients.
z^t is a row vector of the Fisher transformed validity coefficients.
M is a matrix computed as described below.

$$M = \sum^{-1} - \frac{\sum^{-1} UU' \sum^{-1}}{U'\sum^{-1} U}$$

where **U** is a column vector of n 1's where n is the number of validity coefficients.
U^t is a row vector of n 1's where n is the number of validity coefficients.

The null hypothesis is vc = MEM; *i.e.*, that all validity coefficients are equal to the mean effect magnitude. **Q**$_T$, under the null condition, has a chi-square distribution with k-1 degrees of freedom where k is the number of validity coefficients. Rejection of the null hypothesis suggests that the variance in the data set is significantly greater than would be expected if all validity coefficients come from the same population. If the null hypotheses is rejected, the influence of study features on the validity of student ratings is explored. When **Q**$_T$ is calculated for a subset of validity coefficients, partitioned on the basis of a study feature, it becomes **Q**$_W$, the within- category homogeneity statistic and is a test of model specification. If, the null hypothesis, is again rejected, the subset must be resubdivided on the basis of additional study features.

General Linear Model: Testing Influence of Study Features

This general linear model has been described in a number of articles and books

(Becker, 1992; Gleser, & Olkin, 1994; Hedges, & Olkin, 1985; Raudenbush, Becker, & Kalaian, 1988;). The general linear model (Becker, 1994, p. 369) takes the form

$$Y = X\beta + e$$

where **Y** is the vector of validity coefficients;
X is the predictor matrix;
β is the vector of correlation coefficients;
e is a vector of errors.

When data is independent, a weighted least-squares regression can be carried out to test the linear model using SPSS (SPSS, 1994). The weight in question is the inverse of the sampling variance for each study. However, if the data is not-independent, the weight is not the sampling variance, but rather the variance-covariance matrix discussed above.

Construction of predictor matrix. The first task is to construct a predictor matrix (**X**), which specifies the model or models to be tested (Becker, 1992; Raudenbush, Becker, & Kalaian, 1988). In this thesis, a number of models will be tested. When the model is that a given continuous variable explains the variability in a set of data, the predictor matrix (**X**), will consist of a column vector of k 1's (where k is the number of validity coefficients) and a vector of k x's (where x is the value of the study feature associated with each validity coefficient). Mean substitution will be used if less than 5% of the values are missing. Two regression weights will be estimated, the regression weight representing the intercept, and the regression weight representing the slope of the straight line. If more than 5% of the values are missing, the variable will be converted to a categorical variable. Categorical study features will be dummy coded using p-1 columns (where p is the number of levels of the variable). Missing values will be coded 0 in all vectors. Thus. the predictor matrix (**X**), will consist of p columns and a regression weight will be estimated for each value of the variable. Sets of variables can be included in the model by adding column vectors representing each variable to the predictor matrix

.

Calculation of population parameters. The vector of regression coefficients. β, is calculated using the following formula adapted from Hedges and Olkin (1985) and Becker (1994, p. 368):

$$\beta = (X^{t}\sum{}^{-1}X)^{-1}\ X^{t} \sum{}^{-1}\ Z$$

where β is the vector of regression coefficient;
$\mathbf{X}$ is the predictor matrix representing the predictors in a multiple regression model;
$\mathbf{X}^{t}$ is the transposed predictor matrix ;
Σ^{-1} is the inverted variance-covariance matrix;
$\mathbf{z}$ is a column vector of the Fisher transformed validity coefficients.

This vector of regression coefficients is interpreted in the following manner: The regression coefficient associated with the vector of 1's (i.e.. for the intercept) , is the mean effect magnitude or validity coefficient when information on study features is ignored. Each subsequent regression coefficient is the increment in mean effect magnitude associated with information on each predictors in the model

. The standard errors about the regression coefficients were calculated from **V**. the matrix of standard errors, using the formula for **V** adapted from Hedges and Olkin (1985, p. 239) and Becker and Schram (1994, p. 368):

$$V = (X^t \sum^{-1} X)^{-1}$$

where **V** is the variance covariance matrix for the regression coefficients;
X is the predictor matrix for the multiple regression model;
X^t is the transposed predictor matrix;
Σ^{-1} is the inverted variance-covariance matrix for the validity coefficients.

The 95% limits of any linear combination of regression coefficients was calculated using the following formula from Gleser and Olkin (1994, p. 345):

$$95\%CI = \boldsymbol{a}'\beta \pm z_{\alpha/2}\sqrt{(\boldsymbol{a}'Cov(\beta)\boldsymbol{a})}$$

where 95% CI is the 95% confidence interval;
$z_{\alpha/2}$ is 1.96, the two-tailed critical value of the normal distribution;
a is a vector of weights (1);
β is the vector of regression coefficients;
Cov (β) is the vector of variances associated with the regression coefficients.

Testing the model. The overall test for the significance of prediction, is the $\mathbf{Q_R}$ test, given by the equation from Becker and Schram (1994, p. 369):

$$Q_R = \beta'(X^t \sum^{-1} X)\ \beta$$

where $\mathbf{Q_R}$ is the overall test of the predictors;
X is the predictor matrix for multiple regression model;
X^t is the transposed predictor matrix;
β is the vector of regression coefficient;
Σ^{-1} is the inverted variance-covariance matrix for the regression coefficients.

The null hypothesis is $\beta = 0$; *i.e.*, that all regression coefficients are equal to zero. $\mathbf{Q_R}$, under the null condition, has a chi-square distribution with p-1 degrees of freedom where p is the number of predictors (columns) in the predictor matrix **X**. Rejection of the null hypothesis suggests that at least one regression coefficient is non-zero.

Individual regression coefficients can also be tested via the Z test described by

Raudenbush, Becker. and Kalaian (1988. p. 115).

$$Z = \frac{\beta_i}{\sigma_{ii}}$$

where β_i is the i^{th} regression coefficient in the vector of regression coefficients (β);
σ_{ii} is the square root of the i^{th} element in **V**, the variance covariance matrix for the regression coefficients.

The probability values, associated with the test were adjusted using Bonferroni's adjustment. That is, the computed **Z** is compared to the z value for the 100(1 - α/p) the percentile of the standard normal distribution, where p is the number of predictors in the model. For example, if the model being tested has four predictors. the z value computed for each regression coefficient is compared to the z value at the 98.75 percentile (one-tailed test).

The Goodness of Fit ($\mathbf{Q_E}$) test of the regression model is a test that the model adequately explains the variance in the sample data, *i.e.*, that the residual error is not significant. It is calculated by the equation described by Raudenbush. Becker. and Kalaian (1988. p. 115) .

$$Q_E = z^t \sum{}^{-1} z - Q_R$$

where $\mathbf{Q_E}$ is the Goodness of Fit statistic:
$\mathbf{z^t}$ is the transposed vector of validity coefficients:
z is the vector of validity coefficients;
Σ^{-1} is the inverted variance-covariance matrix for the validity coefficients;
$\mathbf{Q_R}$ is the overall test of the predictors.

The null hypothesis is $\delta = \mathbf{X}\beta$ and has k - p degrees of freedom., where k is the number of validity coefficients in the data set, and p is the number of predictors in the model. Rejection of the null hypothesis implies that the model is misspecified and has not taken some significant predictors into account. Normally, we should "fail to reject" this null hypothesis before we conduct the tests for significant predictors described above (Gleser, & Olkin, 1994; Raudenbush, Becker, & Kalaian, 1988).

In conclusion, I conducted three studies. In the first study, I determined whether there were significant and important interactions between the three first order factors, *presenting material*, *facilitating interactions*, and *evaluating learning*; and the study features. If there were, I would subsequently code the outcomes into these three factors. If there were not, I would code the outcomes into the first second order hierarchical factor, *general instructional skill*. I initially constructed a complete model that included the first order factor structure, the study feature in question, and the two way interaction and calculated $\mathbf{Q_{RC}}$ (the Sums of Squares associated with the model). I subsequently constructed a model that only included the two main effects (factor structure and the study feature in question) and calculated $\mathbf{Q_{RM}}$ (the Sums of Squares associated with the main effects). The difference between the two ($\mathbf{Q_{RC}}$ - $\mathbf{Q_{RM}}$) was $\mathbf{Q_{RI}}$ (Sums of Squares associated with the interaction). If the null hypothesis is rejected, the interaction between factor structure and the study feature in question is significant. The presence of a large number of interactions between the factor structure and the study features would indicate that I must use the first order factor structure to code the outcomes since too much information would be lost if I collapsed them into *general instructional skill*.

In the second study, I computed the mean effect magnitude (validity coefficient) for the entire data set, modelling the dependencies among the data. Subsequently I tested the homogeneity of the data set. If the null hypothesis, that the reported validity coefficients were not significantly different from the mean effect magnitude, is rejected, I would conduct a third study.

In the third study, I determined which conditions significantly influenced the validity of student ratings of instruction. For each set of study features *(i.e.*, methodological and publication, quality of evaluation, rating form, achievement criterion, instructor and student explanatory, and course and institutional explanatory), I tested a number of hierarchical models. I subsequently selected significant and practically important study features and constructed a hierarchical model to account for the significant variability in the data set.

Results

*Characteristics of Multisection Validity Studie*s

Forty-three studies were identified that met the inclusion criteria. These are presented in Table 16. Seven hundred and forty one validity coefficients were extracted. Although four studies provided one validity coefficient each; most studies provided more than one validity coefficient (*e.g.,*Centra, 1977, and McKeachie, Linn, & Mann, 1971, each provided 108 validity coefficients). Thus, the interdependencies in the data set were modeled in Study 1.

Table 16. *Characteristics of the 43 studies included in validity set.*

TRF Names	Validity study	n of vc extracted	n of vc analyzed	n of independent sets
Purdue	Bendig (1953)	2	2	1
SIR. ISPI	Benton. & Scott (1976)	34	34	3
TCE	Bolton. Bonge. & Marr (1979)	12	10	1
in house	Braskamp. Caulley. & Costin (1979)	16	16	1
in house	Bryson (1974)	13	12	1
SIR	Centra (1977)	108	72	7
SIRS	Cohen. & Berger (1970)	12	12	1
in house	Costin (1978)	4	4	4
in house	Crooks. & Smock (1974)	8	8	1
SOS	Doyle. & Crichton (1978)	14	12	1
SOS	Doyle. & Whitely (1974)	12	10	1
Purdue	Elliott (1950)	48	40	1
CEQ	Ellis. & Rickard (1977)	4	4	1
in house	Endo. & Della-Piana (1976)	14	14	1
Endevor	Frey (1973)	12	10	1
Endevor	Frey (1976)	21	12	2
Endevor	Frey. Leonard. & Beatty (1975)	14	15	3
in house	Grush. & Costin (1975)	1	1	1
SIR. SECTB.	Hazleton (1980)	38	36	1
CLIC	Hoffman (1978)	44	44	3
in house	Marsh. Fleiner. & Thomas (1975)	10	8	1
SEEQ	Marsh. & Overall (1980)	18	16	1
in house	McKeachie. Lin. , & Mann (1971)	108	118	11
in house	McKeachie. Lin. & Mendelson (1978)	1	1	1
in house	Mintzes (1977)	20	16	1
in house	Morgan. & Vasche (1978)	5	5	1
in house	Morsh. Burgess. & Smith (1956)	10	1	1
in house	Murdock (1969)	1	1	1
in house	Murray (1983)	4	2	1
SOST	Orpen (1980)	7	6	1
in house	Palmer. Carliner . & Romer (1978)	2	1	1
in house	Rankin. Greenmun. & Tracy (1965)	12	1	1
Purdue	Remmers. Martin. & Elliot (1949)	8	8	1
in house	Reynolds. & Hansvick (1978)	4	4	2
in house	Rodin. & Rodin (1972)	1	1	1
in house	Rubinstein. & Mitchell (1970)	4	1	1
in house	Solomon. Roseberg. & Bezdek (1964)	30	28	1
in house	Soper (1973)	2	2	1
in house	Sorge. & Kline (1973)	12	12	1
in house	Sullivan. & Skanes (1974)	14	14	14
in house	Turner. & Thompson (1974)	16	16	2
in house	Wherry (1951)	7	7	4
in house	Whitely. & Doyle (1979)	4	4	2

note vc is validity coefficient

The first question I addressed in this meta-analysis was *Are there any significant and important interactions between the first order factor structure and study features?* The first order factor structure consists of three correlated factors, *presenting material* (Factor 1), *facilitating interactions* (Factor 2), and *evaluating learning* (Factor 4), and a fourth factor (Factor 3), representing miscellaneous behaviours. Therefore, I only considered interactions between the study features and factors 1, 2, and 4. In addition, since the sample sizes are very large, the $Q_R$test is too powerful and trivial interactions (explaining only small amounts of variability) are significant. Therefore, I only explored interactions explaining more than 1% of the variance.

Methodological and Publication Features

The predictive power of the interactions between the first order factor structure and the methodological and publication study features, such as *year of publication*, *number of sections*, *computational issues*, and *study source*, are presented in Table 17. The only significant interaction which explained more than 1% of the variance was the interaction between the first order factor structure and *study source*. *Post-hoc* tests indicate that the validity of student ratings reported in theses when students evaluated their instructor's ability to facilitate interactions (Factor 2) is significantly lower (validity = -.53) from that reported in theses for students evaluating their instructor's ability to either present material (Factor 1) or evaluate learning (Factor 4) (validity = .21 and .22, respectively).

Table 17. *Predictive power of the interactions between the first-order factor structure and methodological and publication study features.*

FEATURE	Q_{RI}	df	sig	%exp
Year of publication	11.99	3	0.001	0.63
Number of sections	4.68	3	ns	0.25
Computational Issues	5.26	6	ns	0.28
Study Source	47.83	6	0.001	2.52

Note Q_{RI} is the Sum of Squares associated with the interactions. df is the degrees of freedom. sig is the significance level. and % exp is the percent of the variability in the data set explained by the interactions.

Quality of Evaluation Study Features

The predictive power of the interactions between the first order factor structure and the quality of evaluation study features, such as *timing*, *administrator*, *scoring bias*, *test bias*, and *group equivalence*, are presented in Table 18. There are no significant interactions between the first order factor structure and the quality of evaluation study features that explain more than 1% of the variance.

Table 18. *Predictive power of the interactions between the first-order factor structure and quality of evaluation study features.*

FEATURE	Q_{RI}	df	sig	% exp
Timing	10.30	6	ns	0.52
Administrator	11.30	5	0.05	0.59
Scoring Bias	4.97	6	ns	0.26
Test Bias	10.82	6	ns	0.57
Group Equivalence	10.11	9	ns	0.53

Note Q_{RI} is the Sum of Squares associated with the interactions. df is the degrees of freedom. sig is the significance level. and % exp is the percent of the variability in the data set explained by the inbteractions.

Student Rating Form Study Features

The predictive power of interactions between the first order factor structure and student rating form study features, such as *source*, *response scale*, *anonymity*, *length*, *reliability*, *factor length*, *completion rate*, and *diversity index*, are presented in Table 19. The only significant interaction which explains more than 1% of the variance was the interaction between the first order factor structure and *source*. However, *post-hoc* tests indicate that there are no significant differences in validity across presenting material (Factor 1), facilitating interactions (Factor 2) and evaluating learning (Factor 4).

Table 19. *Predictive power of interactions between the first-order factor structure and student rating form study features*

FEATURE	Qr	df	sig	%exp
Source	106.7	6	0.001	5.6
Response Scale	14.8	8	ns	0.78
Anonymity	0.73	3	ns	0.04
Length	4.31	3	ns	0.67
Reliability	10.4	6	ns	0.55
Factor length	6.39	3	ns	0.34
Completion rate	4.35	6	ns	0.23
Diversity Index	15.79	3	0.01	0.83

Note Q_{RI} is the Sum of Squares associated with the interactions. df is the degrees of freedom, sig is the significance level, and % exp is the percent of the variability in the data set explained by the inbteractions.

Achievement Measure Study Features

The predictive power of interactions between first order factor structure and achievement measure study features, such as *source*, *type*, *number*, *length*, *source calibration*, *learning criteria*, *scale*, *value*, and *reliability*, are presented in Table 20. There are three significant interactions with the first order factor structure which explain more than 1% of the variance: *source, achievement test type,* and *value*. However, *post-hoc* tests again indicate that there are no significant differences in validity across the three factors (presenting material, facilitating interactions, aud evaluating learning .

Table 20. *Predictive power of interactions between first-order factor structure and achievement measure study features.*

FEATURE	Q_{RI}	df	sig	%exp
Source	20.89	6	0.01	1.10
Type	20.21	9	0.02	1.06
Number	18.17	9	0.05	0.96
Length	12.23	6	ns	0.64
Score calibration	8.87	9	ns	0.47
Learning Criteria	16.06	9	ns	0.85
Scale	2.28	3	ns	0.12
Value	29.11	9	0.001	1.53
Reliability	18.73	6	0.01	0.99

Note Q_{RI} is the Sum of Squares associated with the interactions. df is the degrees of freedom, sig is the significance level, and % exp is the percent of the variability in the data set explained by the inbteractions.

Instructor and Student Explanatory Study Features

The predictive power of the interactions between first order factor structure and instructor/ student explanatory characteristics, such as *instructor rank, instructor experience, instructor autonomy,* and *student gender*, are presented in Table 21. There were two significant interactions with the first order factor structure which explained more than 1% of the variance: *instructor autonomy* and *student gender*. *Post-hoc* tests indicated that the validity of student ratings was significantly lower (validity = 0) when students rated instructors with a high degee of autonomy on their ability to facilitate interactions (Factor 2) than it was when students rated instructors with a high degree of autonomy on their ability to either present material (Factor 1) or evaluate learning (Factor 4), validity= .41 and .36, respectively. *Post-hoc* tests indicated that the validity of student ratings was significantly lower (validity = .01) when male students rated the instructor's ability to present material (Factor 1) compared to their rating of their instructor's ability to facilitate interactions (Factor 2) or evaluate learning (Factor 4), validity = .22 and .23, respectively. However, the validity of student ratings was significantly lower when female students rated their instructor's ability to facilitate interactions (Factor 2) (validity = .01) than it was when they either rated their instructor's ability to present materials (Factor 1) or evaluate learning (Factor 4), validity = .26 and .26, respectively.

Table 21. *Predictive power of interactions between first-order factor structure and instructor and student explanatory study features*

FEATURE	Q_{RI}	df	sig	%exp
Rank	11.27	6	ns	0.59
Experience	12.69	6	0.05	0.67
Autonomy	38.88	9	0.001	2.05
Gender	27.39	6	0.001	1.44

Note Q_{RI} is the Sum of Squares associated with the interactions. df is the degrees of freedom, sig is the significance level, and % exp is the percent of the variability in the data set explained by the inbteractions.

Course and Institutional Explanatory Study Features

The predictive power of the interactions between first order factor structure and course and institutional explanatory characteristics, such as *type of instruction, teaching duration, discipline, course length, season, type of institution, class size,* and *section size*. are presented in Table 22 . There were three significant interactions with the first order factor structure which explained more than 1% of the variance: *type of instruction, season,* and *section size*. *Post-hoc* tests indicated that the validity of student ratings was significantly lower (validity = -.14) when students in lectures rated their instructor's ability to facilitate interactions (Factor 2) than it is when students in lectures rated their instructor's ability to either present material (Factor 1) or evaluate learning (Factor 4), validity = .16 and .20, respectively; or, when students in discussion/tutorial classes rated their instructor's ability to either present material (Factor 1) or facilitate interactions (Factor 2), validity = .28 and .22, respectively. P*ost-hoc* tests indicated that there were no significant differences in validity with either season or section size across presenting material (Factor 1), facilitating interactions (Factor 2), and evaluating learning (Factor 4).

Table 22. *Predictive power of interactions between first-order factor structure and course and institutional explanatory study features*

FEATURE	Q_{RI}	df	sig	%exp
Instruction	25.39	11	0.01	1.34
Teaching duration	16.68	6	0.02	0.87
Discipline	1.25	3	ns	0.17
Course length	18.92	5	0.001	0.99
Season	19.96	6	0.01	1.05
Type of institution	4.01	2	ns	0.21
Class size	16.43	9	ns	0.87
Section size	30.14	6	0.001	1.59

Note Q_{RI} is the Sum of Squares associated with the interactions. df is the degrees of freedom. sig is the significance level, and % exp is the percent of the variability in the data set explained by the inbteractions.

Thus, out of 38 tests for interactions between the first order factor structure and study features, only four were both significant and practically important. In general, they indicated that under certain conditions (reported in theses, instructors with complete autonomy, female students, lecture instructional format) the validity of student ratings assessing the instructor's facilitation of interactions (Factor 2) was significantly lower than the validity of ratings assessing the instructor's presentation of material (Factor 1) or the evaluation of learning (Factor 4). In addition, the validity of student ratings was not significanly different than 0 when male students assessed the instructor's presentation of material (Factor 2) .

Study 2: Overall Validity of Student Ratings of Instruction

Since there are very few interactions between the first-order structure and the study features, and since Hierarchical Factor 1 represents a composite of first order Factors 1, 2, and 4, I decided to limit my subsequent investigations to Hierarchical Factor 1: *General Instructional Skill*. Thus, I dropped seventy-six validity coefficients because they were calculated from items which did not contribute to this factor. Thus, the final data set consisted of 665 validity coefficients.

The weighted average validity coefficient for the 665 outcomes is .33. The 95% confidence interval (weighted by the number of class sections) extends from .29 to .37.

However, the correlation between student ratings and student learning is attenuated by unreliability in both the rating and achievement instruments. The average reliabilities of the student rating and achievement instruments in the multisection validity studies are .74 and .69, respectively. Therefore, when the correlation coefficient between student ratings of *general instructional skill* and student learning is corrected for attenuation (Downie, & Heathe, 1974), the correlation becomes .47 with a 95% confidence interval extending from .43 to .51.

Thus, there is a moderate association between student ratings and student learning. According to Cohen (1988, p 80) , a correlation of this size "would be perceptible to the naked eye by a reasonably sensitive observer". Thus, student ratings of instructional effectiveness are reasonably valid. The homogeneity statistic, Q_T, is 1500.3 indicating that the data set is heterogeneous and that study features may moderate the validity of student ratings. If I had ignored non-independence, the Q_T would have been only 1120.3. Thus, it appears that ignoring non-independence and treating the data as if it were independent reduces the homogeneity statistic, and makes it easier to detect homogeneous sets.

Study 3: Influence of Study Features

For each subset of study features (Methodological and Publication, Quality of Evaluation, Student Rating Form, Achievement Measure, Instructor and Student Characteristics, and Course and Institutional Characteristics) the results of individual study features are described first, followed by the results of the multiple regression models for various combinations of study features within each category. Subsequently, several hierarchical multiple regression models for the complete data set are tested.

Methodological and Publication Features

The influence of methodological and publication study features, such as *year of publication*, *number of sections*, *computational issues*, and *study source*, for the 665 validity coefficients are presented in Table 23. Both *computational issues* and *study source* are significant predictors, predicting 2.7% and 2.4% of the variability in validity coefficients, respectively. Thus the statistic used to represent the association between student ratings of instruction and student achievement and the source of the primary study significantly influenced the validity of student ratings of instruction.

Table 23. *Predictive power of methodological and publication study features.*

FEATURE	Q_R	df	sig	%exp
Year of publication	.23	1	ns	.01
Number of sections	.08	1	ns	.01
Computational Issues	4.48	2	.005	2.70
Study Source	44.44	2	.005	2.96

Note Q_R is the Sum of Squares associated with the predictors. df is the degrees of freedom. sig is the significance level. and % exp is the percent of the variability in the data set explained by the predictors.

Computational issues. Table 24 indicates that when studies report rank correlations, the average validity coefficient is 0.54. The within class fit statistic, Q_w, of 18.9 (df = 18) indicates that the set is homogeneous. Thus, it appears that there is a consistent overestimation of validity coefficients when rank correlations are converted to

Pearson product moment correlations. On the other hand, the average validity coefficients for reported averages and for Pearson products are 0.20 and 0.33, respectively. In neither case is the subset homogeneous. The *z* tests indicate that the validity coefficients computed from rank correlations (MEM = .54) are significantly different from those computed from averages (MEM = .22), but not from those computed from Pearsons's product coefficients (MEM = .33).

Table 24. *Predictors in regression model with methodological and publication study features.*

FEATURE	LEVEL	k	MEM	SE	z test	Q_w
Year of publication	intercept	665	.38	.10	3.91	
	slope		-.01	.02	-.52	
Number of sections	intercept	665	.32	.04	9.493	
	slope		.00	.00	.29	
Computational Issues	from rank	19	.54	.15	3.86	18.8*
	from averages	69	.20	.22	-2.54	106.0
	direct (r)	577	.33	.21	-1.63	1316.7
Study Source	theses	36	-.28	.13	-2.23	60.7
	published reports	114	.25	.18	4.04	249.5
	published articles	515	.38	.18	5.30	1097.3

note k is the number of validity coefficients, MEM is the mean effect magnitude backtransformed to r, SE is the standard error about the mean, Q_w is the within-group homogeneity statistic, and * indicates that the set is homogeneous at α=.05.

Study source. Table 24 also indicates that the average validity coefficients computed for findings extracted from theses, unpublished reports, and published studies, are -0.28, 0.25, and 0.38, respectively. The *z*-tests indicate that the average validity coefficients computed both for reports and published studies differ significantly from that computed for theses. However, the homogeneity tests indicate that none of the subsets are homogeneous.

Multiple regression models of methodological and publication features. Three multiple regression models were tested and are presented in Table 25. As can be seen, adding both computational issues and study source (Model # 2), significantly adds to the prediction compared to only adding computational issues (Model # 1). However, there is no significant increase when additional methodological and publication study features are added (Model # 3). Therefore, only computational issues and study source will be retained in the final multiple model. These two methodological and publication features explained 4.9 % of the variability in validity coefficients.

Table 25. *Multiple regression models of methodological and publication study features.*

#	Model Description	%exp	Q_r	df	Δ Q_r	Δ df
1	Computational Issues	3.1	312.7	2		
2	Computational Issues, Study Source	4.8	345.7	4	30.0 *	2 (1)
3	All predictors	5.1	348.6	6	2.9	2 (2)

note % exp is the cumulative percent explained by the model, QE is the Goodness of Fit statistic, df are the degrees of freedom for the model ΔQ_r and Δdf are the differences in the Goodness of Fit statistic and associated degrees of freedom between the indicated model and the previous model.

Quality of Evaluation Study Features

The influence of study features reflecting the quality of the evaluation, such as *timing, administrator, scoring bias, test bias,* and *group equivalence*, for the 665 validity coefficients are presented in Table 26. Four of the five quality of evaluation characteristics, *i.e*., *timing, scoring bias, test bias,* and *group equivalence* are significant predictors, predicting 2.6%, 2.8%, 2.0%, and 1.6% of the variability in validity coefficient respectively. Thus, characteristics affecting the quality of the evaluation, such as the timing of the evaluation, who evaluates the students' achievement on the common coefficient respectively. Thus, characteristics affecting the quality of the evaluation, such as the timing of the evaluation, who evaluates the students' achievement on the common final, whether the instructor has knowledge of the common final during instruction, and the academic equivalence of sections, predicteu the size of the validity coefficients.

Table 26. *Predictive power of quality of evaluation characteristics*

FEATURE	Q_R	df	sig	% exp
Timing	38.30	2	.005	2.55
Administrator	.58	2	ns	.04
Scoring Bias	42.01	2	.005	2.80
Test Bias	30.47	2	.005	2.03
Group Equivalence	23.99	4	.005	1.60

Note Q_R is the Sum of Squares associated with the predictors. df is the degrees of freedom. sig is the significance level. and % exp is the percent of the variability explained by the predictors.

Timing. Table 27 indicates that the average validity coefficients are 0.38 and 0.50 when the instructional evaluation is carried out after and before the final examination, respectively. The average validity coefficient is 0.21 if the timing of the instructional evaluation is not reported. The z-tests indicate that the validity coefficients, when timing is reported, are significantly different from that when timing is not reported. Homogeneity tests indicate that only the subset of studies which carried out instructional evaluation after the final examination is homogeneous (Q_w = 73.6 (df = 75)).

Scoring bias. Table 27 also indicates that average validity coefficients are 0.18 and 0.45 when the final examination was scored by the instructor and by a person other than the instructor, respectively. The average validity coefficient is 0.22 if the study did not indicate who scored the final examination. The z-tests indicate that the average validity coefficient when the final examination was scored by someone other than the instructor is significantly greater than when either the scorer was the instructor or when the scorer was not reported. Homogeneity tests indicate that only the subset of studies in which the instructor scored the final examination is homogeneous ($Q_w = 21.9$ (df = 33)).

Table 27. *Predictors in regression model with quality of evaluation study features.*

FEATURE	LEVEL	k	MEM	SE	z test	Q_w
Timing	unknown	302	.21	.03	7.13	901.9
	after exams	76	.50	.07	5.45	73.6*
	before exams	287	.38	.05	4.95	492.4
Administrator	unknown	389	.32	.03	12.32	869.2
	instructor	20	.40	.13	.72	21.7*
	other	256	.34	.05	.36	608.7
Scoring Bias	unknown	366	..22	.04	7.48	825.1
	instructor	34	.18	.06	-.52	21.9*
	researcher/ other	265	.45	.06	6.17	611.3
Test Bias	unknown	425	.26	.03	9.55	1118.1
	prior knowledge	76	.22	.07	-.66	63.0*
	no prior knowledge	164	.46	.05	5.13	288.9
Group Equivalence	unknown	283	.28	.02	11.52	556.8
	stated equivalence	16	.55	.13	2.53	13.4*
	statistical control	333	.32	.03	3.35	793.8
	experimental control	33	.46	.07	3.32	28.3*

note k is the number of validity coefficients. MEM is the mean effect magnitude backtransformed to r. SE is the standard error about the mea. Q_w. * indicates that the set is homogeneous at α=.05.

Test bias. Table 27 also indicates that average validity coefficients were 0.22 when the instructor had prior knowledge of the final examination and 0.46 when the instructor had no prior knowledge of the final examination. The average validity coefficient is 0.26 if the study did not report whether or not the instructor had prior knowledge of the final examination. The z-tests indicate that the average validity coefficient, when the instructor had no prior knowledge is significantly higher than that when either the instructor had prior knowledge or when prior knowledge is not reported. Homogeneity tests indicate that only the subset of studies which reported that the instructor had no prior knowledge of the final examination is homogeneous ($Q_w = 63.0$ (df = 75)).

Group equivalence. Table 27 also indicates that average validity coefficients were 0.28 when there was no information on whether the students had equal ability across sections or when the sections were not equivalent, 0.32 when statistical control was used to control for non-equivalence, 0.46 when random assignment was used to experimentally control for non-equivalence, and 0.55 when the primary investigators reported that sections were equivalent. The z-tests indicate that the average validity coefficients for the three subsets with some form of group equivalence are significantly higher than when groups were inequivalent or when there was no information on group equivalence. Homogeneity tests indicate that two subsets of studies are homogeneous: studies which reported that the sections were equivalent (($Q_w = 13.4$ (df = 15)) and studies which used random assignment ($Q_w = 28.30$ (df = 32)).

Multiple regression models of quality of evaluation features. Five multiple regression models were tested and are presented in Table 28. As can be seen, adding both scoring bias and timing (Model # 2), significantly adds to the prediction compared to only adding scoring bias (Model # 1). Likewise, adding group equivalence (Model #3) to scoring bias and timing (Model # 2) significantly adds to the prediction. However, there is no significant increase when additional quality of evaluation study features (Models # 4 and 5) are added. Therefore, only scoring bias, timing, and group equivalence will be retained in the final multiple model. These three quality of evaluation study features explain 5.9% of the variability in validity coefficients.

Table 28. *Multiple regression models of quality of evaluation study features*

#	Model Description	% esp	Q_r	df	Δ Q_r	Δ df
1	Scoring Bias	2.80	314.2	2		
2	Scoring Bias. Timing	5.06	348.1	4	33.9 *	2 (1)
3	Scoring Bias. Timing. Group Equivalence	5.86	360.0	7	11.9 *	3 (2)
4	Scoring Bias. Timing, Group Equivalence. Test Bias	6.02	362.5	9	2.5	2 (3)
5	All predictors	6.31	366.8	11	4.3	2 (4)

note % exp is the cumulative percent explained by the model, QE is the Goodness of Fit statistic, df are the degrees of freedom for the model ΔQ,and Δdf are the differences in the Goodness of Fit statistic and associated degrees of freedom between the indicated model and the previous model.

Student Rating Forms Study Features

The influence of characteristics reflecting student rating forms. such as *source, response scale, student anonymity, length. reliability, factor length. completion rate, diversity index*, and *structure* for the 665 validity coefficients are presented in Table 29. Six of nine study features, *i.e., response scale, length, factor length, completion rate, diversity index*, and *structure* are significant predictors, predicting 0.9%. 1.9%, 0.5%, 1.7%, 0.6%, and 8.9% of the variability in validity coefficient respectively. Thus, properties of the student rating forms, such as the number of choices on the form, number of items on the form, number of items contributing to the validity coefficient, proportion of students within the class completing the evaluation. diversity of the items contributing to the validity coefficient, and dimensional structure of the items contributing to the validity coefficient, significantly influenced the validity of student ratings of instruction.

Table 29. *Predictive power of student rating form study features*

FEATURE	Q_R	df	sig	%exp
Source	.34	2	ns	.02
Response Scale	13.87	3	.005	.93
Anonymity	.30	1	ns	.02
Length	38.29	1	.005	1.88
Reliability	.68	2	ns	.05
Factor length	7.42	1	.010	.49
Completion rate	26.030	2	.005	1.74
Diversity Index	8.87	1	.005	.59
Structure	133.9	4	.005	8.93

Note Q_R is the Sum of Squares associated with the predictors. df is the degrees of freedom, sig is the significance level. and % exp is the percent of the variability in the data set explained by the predictors.

Length. Table 30 indicates that the length of the student rating form is negatively correlated to validity coefficient. That is. short forms had higher validity coefficients than did long forms. That is. all other things being equal. a rating form of ten items will have a validity coefficient of 0.37; while a rating form of thirty items will have a validity coefficient of 0.17.

Response scale. Table 30 also indicates that the average validity coefficient is 0.31 when the response scale of the student rating form was not reported, 0.30 when the response scale was a forced choice, 0.32 when the response scale was a Likert scale using

Table 30. *Predictors in regression model with student rating form study features*

FEATURE	LEVEL	k	M	se	z test	Q_W
Source	local	247	.34	.03	12.8	391.4
	departmental	281	.32	.05	-.52	790.0
	standardized	137	.33	.05	-.17	291.6
Response Scale	unknown	253	.31	.03	9.66	505.0
	forced choice	14	.30	.14	-.10	24.6
	2 tc 5	316	.32	.05	.24	876.8
	6 to 25	82	.59	.10	3.64	80.0*
Anonymity	unknown/no	593	.34	.02	15.6	1433.0
	yes	72	.30	.07	-.54	66.9*
Length	intercept	665	.47	.03	16.1	
	slope		-.01	.00	-6.19	
Reliability	unknown	378	.34	.03	12.5	710.5
	<71.5	151	.31	.05	-.73	422.5
	>71.5	136	.33	.05	-.38	412.1
Factor length	intercept	665	.37	.02	16.4	
	slope		.00	.00	-2.72	
Completion rate	unknown	536	.28	.02	11.6	1256.4
	<80	74	.52	.07	4.38	89.0
	>80	55	.44	.06	3.25	128.9
Diversity Index	intercept	665	.38	.02	16.0	
	slope		-.02	.01	-2.98	
Structure	intercept	665	.18	.03	7.07	
	Factor 1 slope		.13	.02	7.33	
	Factor 2 slope		.13	.02	7.43	
	Factor 3 slope		.05	.03	1.92	
	Factor 4 slope		.19	.03	6.77	

note k is the number of validity coefficients. MEM is the mean effect magnitude backtransformed to r. SE is the standard error about the mea. Q_W. * indicates that the set is homogeneous at α=.05.

2 to 5 choices, and 0.59 when the response scale was a Likert scale using 6 or more choices. The z-tests indicate that the average validity coefficient when the response scale was a Likert scale using 6 or more choices is significantly higher than the average validity coefficients for the other subgroups. Moreover, the homogeneity test indicates that only the subset of studies which reported that the response scale was a Likert scale

using 6 or more choices is homogeneous ($Q_w = 80.0$ (df = 81)).

Factor length. Likewise, Table 30 indicates that the number of items included in the specific factor was negatively correlated to the validity coefficient. That is, factors assessed with only a few items had higher validity coefficients than did longer factors. For example, all other things being equal, a factor of three items will have a validity coefficient of 0.36; while a factor of ten items will have a validity coefficient of 0.33.

Completion rate. Table 30 also indicates that the average validity coefficient is 0.28 when the completion rate was not reported, 0.52 when the completion rate was less than 80%, and 0.44 when it was over 80%. The z-tests indicate that when the completion rate was reported, the average validity coefficient is significantly different from that when the completion rate was not reported. Homogeneity tests indicate that none of the subsets are homogeneous.

Diversity index. Table 30 also indicates that the diversity of items contributing to the validity coefficient is negatively correlated to validity. That is, the greater the diversity of items contributing to the validity coefficient, the lower the validity of student ratings of instruction. For example, when the student rating was unidimensional (diversity index = 0), the validity coefficient is 0.38. However, when the student rating form was multidimensional (*e.g.*, *Skill* has a diversity index of 4.0 in the McKeachie, Lin, and Mendelson, 1978 study) the validity coefficient is 0.30.

Structure. Table 30 also indicates that the dimensional structure of the student rating form significantly influences validity. The z-tests indicate that the average validity coefficient is 0.18 when factor structure is ignored, but validity increased significantly as the proportion of items in factor 1, factor 2, or factor 4 increased. However, an increase in the proportion of items in factor 3 did not increase the size of the validity coefficient. That is, the addition of items assessing the instructor's *Discipline*, *Choice of Required Materials*, *Knowledge of Domain* and, *Use of Objectives* did not increase the validity coefficient over that for items assessing the instructor's role as delivering information, facilitating a learning environment, and evaluating learning.

Multiple regression models of student rating form study features. Six multiple regression models were tested and are presented in Table 31. As can be seen, adding both structure and length (Model # 2), significantly adds to the prediction compared to only adding structure (Model # 1). Likewise, adding completion rate (Model #3) to structure and length (Model # 2) significantly adds to the prediction. However, there is no significant increase when additional student rating form study features (Models # 4, 5, and 6) are added. Therefore, only structure, completion rate, and length will be retained in the final multiple model. These three student rating study features explained 12.8% of the variability in validity coefficients.

Table 31. *Multiple regression models of student rating form study features*

#	Model Description	%exp	Q_r	df	Δ Q_r	Δ df
1	Structure	8.9	406.1	4		
2	Structure, Length	10.7	433.2	5	27.1 *	1 (1)
3	Structure, Completion Rate, Length	12.8	463.8	7	30.6 *	2 (2)
4	Structure, Completion Rate, Length, Response	13.2	469.6	10	4.8	3 (3)
5	all statisticall significant predictors	13.6	475.9	14	6.3	4 (4)
6	All predictors	13.8	478.6	17	2.7	3 (5)

note % exp is the cumulative percent explained by the model, QE is the Goodness of Fit statistic, df are the degrees of freedom for the model Δ$Q_r$and Δdf are the differences in the Goodness of Fit statistic and associated degrees of freedom between the indicated model and the previous model.

Achievement Measure Study Features

The influence of achievement study features, such as *source, type, number, length, score calibration, learning criteria, scale, value,* and *reliability*, for the 665 validity coefficients are presented in Table 32. Two of the nine achievement study features, *i.e.*, *type* and *score calibration,* are significant predictors, predicting 2.4% and 1.2% of the

Table 32. *Predictive power of achievement measure study features.*

FEATURE	Q_R	df	sig	%exp
Source	4.65	3	ns	.35
Type	32.46	4	.005	2.42
Number	1.25	4	ns	.09
Length	2.66	3	ns	.20
Score calibration	15.90	4	.005	1.18
Learning Criteria	1.97	4	ns	.15
Scale	4.96	3	ns	.37
Value	6.57	4	ns	.49
Reliability	2.26	3	ns	.17

Note Q_R is the Sum of Squares associated with the predictors, df is the degrees of freedom, sig is the significance level, and % exp is the percent of the variability in the data set explained by the predictors.

variability in validity coefficients, respectively. Thus, characteristics of the final examination used as a criterion measure, such as the type of test used and the manner in which the test score was calibrated, are significant predictors of the size of the validity coefficient.

Type. Table 33 indicates that the average validity coefficient is 0.34 when the test type was unknown, and 0.34, 0.18, and 0.26 when the final examination consisted of problem-solving or skill questions, essay questions, and multiple-choice questions, respectively. The z-tests indicate that the average validity coefficient when the type of test was reported is significantly different from that when the test type was not reported. However, the homogeneity test indicates that only the subset of validity coefficients for essay test is homogeneous ($Q_w = 26.3$ (df = 41)).

Score calibration. Table 33 also indicates that the average validity coefficient is 0.31 when the score calibration of the final achievement test was not reported, and 0.39, 0.19, and 0.40 when raw, weighted, and standardized scores were used respectively. The z-tests indicate that the average validity coefficient when standardized scores were used is significantly different from that calculated when the score calibration of the final test was not reported. Homogeneity tests indicate that none of the subset of studies are homogeneous.

Table 33. *Predictors in regression with achievement measure study features*

FEATURE	LEVEL	k	mem	se	z test	Q_W
Source	unknown/local	212	.37	.03	14.37	528.3
	test bank	396	.31	.03	-2.49	836.3
	standardized	57	.36	.04	-.14	100.3
Type	unknown	306	.43	.03	14.21	609.5
	skill/problems	93	.34	.05	-2.24	246.4
	essay	42	.18	.10	-3.16	26.3*
	multiple-choice	224	.26	.04	-4.71	585.8
Number	unknown	36	.43	.11	4.07	21.9*
	multiple	134	.33	.12	-.95	488.3
	pre- and post-	111	.26	.12	-1.56	382.9
	one	384	.35	.12	-.79	588.8
Length	unknown	499	.34	.02	16.32	1006.7
	< 20	69	.29	.04	-1.79	231.0
	> 20	97	.29	.03	-2.10	227.1
Score calibration	unknown	392	.31	.03	12.06	914.2
	raw	168	.39	.05	1.96	215.7
	weighted	72	.19	.07	-1.75	279.1
	standardized	33	.40	.03	3.08	71.6
Learning Criteria	unknown	358	.36	.03	13.70	651.3
	affective	12	.28	.08	-1.18	13.2*
	general/factual	94	.29	.05	-1.90	215.2
	comprehension	201	.30	.04	-1.73	590.2
Scale	unknown	47	.40	.06	6.88	38.5*
	letter	36	.20	.12	-.09	27.4*
	numerical	582	.33	.09	-.93	1430.3
Value	unknown	492	.31	.03	12.28	948.7
	<40	57	.45	.06	2.55	70.5
	>40 and <100	44	.29	.06	-.48	243.4
	100	72	.37	.10	.63	229.9
Reliability	unknown	535	.34	.02	15.08	1028.9
	<70	78	.21	.07	-2.15	321.7
	>70	52	.44	.08	1.48	142.1

note k is the number of validity coefficients. MEM is the mean effect magnitude backtransformed to r. SE is the standard error about the mean Q_W. * indicates that the set is homogeneous at α=.05.

Multiple regression models of achievement measure study features. Three multiple regression models were tested and are presented in Table 34. Adding both score calibration and test type (Model # 2), significantly added to the prediction compared to adding only test type (Model # 1). However, adding all the achievement measure study features (Model #3) also significantly added to the prediction. This probably reflects the addition of so many (7) individually non-significant predictors, which in combination add significantly to the prediction. Nevertheless, only test type and score calibration were retained in the final multiple model. These two achievement measure study features explained 2.8 % of the variability in validity coefficients.

Table 34. *Multiple regression models of achievement measure study features*

#	Model Description	%exp	Q_r	df	Δ Q_r	Δ df
1	Type	2.4	301.1	3		
2	Type, Score Calibration	2.8	313.9	6	12.8 *	3
3	All predictors	6.1	364.3	23	50.7 *	17

note % exp is the cummultative percent explained by the model. QE is the Goodness of Fit statistic. df are the degrees of freedom for the model Δ$Q_r$and Δdf are the differences in the Goodness of Fit statistic and associated degrees of freedom between the indicated model and the previous model.

Instructor and Student Explanatory Study Features

The influence of explanatory features, such as *instructor rank, instructor experience, instructor autonomy,* and *gender* on the 665 validity coefficients are presented in Table 35. Three of the four explanatory study features, *i.e., instructor rank, instructor experience,* and *instructor autonomy* were significant predictors, predicting 3.2%, 2.3%, and 2.1% of the variability in validity respectively. Thus, instructor explanatory characteristics such as rank, experience and autonomy significantly influence the validity of student ratings of instruction.

Table 35. *Predictive power of instructor and student explanatory study features*

FEATURE	Q_R	df	sig	%exp
Rank	47.32	2	.005	3.15
Experience	34.05	2	.005	2.27
Autonomy	31.41	3	.005	2.09
Gender	2.69	2	ns	.18

Note Q_R is the Sum of Squares associated with the predictors. df is the degrees of freedom, sig is the significance level, and % exp is the percent of the variability in the data set explained by the predictors.

Instructor rank. Table 36 indicates that the average validity coefficients is 0.21 when the instructors' ranks were either not reported or mixed, 0.39 when the instructors were teaching assistants, and 0.54 when the instructors were faculty. The z-tests indicate that the average validity coefficient when studies report the instructors' ranks is significantly different from those in studies which did not differentiate between teaching assistants and faculty. Homogeneity tests indicate that only the subset of validity coefficients computed for faculty is homogeneous (Q_w = 124 (df = 118)).

Instructor experience. Table 36 also indicates that the average validity coefficient is 0.28 when the instructors' experience was not reported or was mixed, and 0.51 and 0.55 when it was reported to be less than one year and more than one year, respectively. The z-tests indicate that the average validity coefficient when studies reported the instructors' experience is significantly different from those in studies which did not differentiate between sections taught by experienced and not experienced instructors. Homogeneity tests indicate that only the subset of validity coefficients computed for experienced instructors is homogeneous (Q_w = 83.9 (df = 74)).

Instructor autonomy. Table 36 also indicates that the average validity coefficient is 0.22 when the instructors' degree of autonomy was not reported, and 0.40, 0.48, and 0.38 when the instructors' autonomy was reported to be low, medium, and high, respectively. The z-tests indicate that the average validity coefficient when studies repored the instructors' autonomy is significantly different from those in studies which did not report the instructors' autonomy. However, the *z*- tests indicate that there are no significant differences among the studies reporting the instructors' autonomy. In addition, homogeneity tests indicate that none of the subsets are homogeneous.

Table 36. *Predictors in regression model with instructor and student explanatory study features.*

FEATURE	LEVEL	k	mem	se	z test	Q_w
Rank	unknown/mixed	314	.21	.03	7.21	784.3
	TA's	232	.39	.05	4.29	544.4
	faculty	119	.54	.06	6.50	124.0*
Experience	unknown/mixed	541	.28	.02	12.50	1244.5
	new	49	.52	.08	3.75	137.8
	experienced	75	.55	.07	4.82	83.9*
Autonomy	unknown	294	.22	.03	7.28	804.6
	directed	99	.40	.06	3.33	258.4
	co-ordinated	142	.48	.06	5.03	172.2
	autonomous	130	.38	.06	3.19	233.7
Student gender	mixed/unknown	330	.36	.03	13.22	624.9
	female	56	.29	.08	-.96	117.1
	male	279	.30	.05	-1.51	755.5

note k is the number of validity coefficients. MEM is the mean effect magnitude backtransformed to r. SE is the standard error about the mea, Q_w is the within-group homogeneity statistic. * indicates that the set is homogeneous at α=.05.

Multiple regression models of instructor and student explanatory features. Four multiple regression models were tested and are presented in Table 37. As can be seen, adding both rank and experience (Model # 2), significantly adds to the prediction compared to only adding rank (Model # 1). Likewise, adding autonomy (Model #3) to rank and experience (Model # 2) significantly adds to the prediction. However, there is no significant increase when additional instructor and student explanatory study features (Model # 4) are added. Therefore, only rank, experience, and autonomy will be retained in the final multiple model. These three variables explain 5.2 % of the variability.

Course and Institutional Study Features

The influence of explanatory features, such as *type of instruction, teaching duration, discipline, course length, season, institution type, class size,* and *section size* on the 665 validity coefficients are presented in Table 38. Four of the eight explanatory study features, *i.e., teaching duration, discipline, course length,* and *season,* are significant predictors, predicting .5%, 1.9%, .7%, and 1.4% of the variability in validity coefficients, respectively. Thus, course explanatory features such as discipline, class length, course duration and semester, are significant predictors of validity.

Table 37. *Multiple regression models of instructor and student explanatory study features.*

#	Model Description	%exp	Q_r	df	Δ Q_r	Δdf
1	Rank	3.2	319.5	2		
2	Rank, Experience	3.8	328.7	4	19.2 *	2 (1)
3	Rank, Experience, Autonomy	5.2	349.7	7	21.0 *	3 (2)
4	All predictors	5.5	354.6	9	4.9	2 (3)

note % exp is the cumulative percent explained by the model. QE is the Goodness of Fit statistic. df are the degrees of freedom for the model $\Delta Q_r$and Δdf are the differences in the Goodness of Fit statistic and associated degrees of freedom between the indicated model and the previous model.

Table 38. *Predictive power of course and institutional explanatory study features.*

FEATURE	Q_R	df	sig	pexp
Instruction	9.85	5	ns	.66
Teaching duration	7.00	2	.05	.47
Discipline	28.31	1	.005	1.89
Course length	10.09	2	.005	.67
Season	20.27	2	.005	1.35
Type of institution	.27	1	ns	.02
Class size	28.6	3	.005	1.9
Section size	18.6	3	.005	1.2

Note Q_R is the Sum of Squares associated with the predictors. df is the degrees of freedom. sig is the significance level. and % exp is the percent of the variability in the data set explained by the predictors.

Teaching duration. Table 39 indicates that the average validity coefficient is 0.32 when the studies did not report on class length, 0.44 when class length was less than 3 hours a week, and 0.26 when the class length was more than 3 hours a week. The z-tests indicate that the average validity coefficient when the class length was less than 3 hours a week is significantly different from that when the teaching duration was unknown. However, homogeneity tests indicate that none of the subsets are homogeneous.

Discipline. Table 39 also indicates that the average validity coefficient is 0..25 when the course discipline was arts, letters, and social science but 0.44 when it was science or mathematics. The z-test indicates that the average validity coefficient for science courses is significantly different from that for arts, letters, and social science courses. However, homogeneity tests indicate that neither subset of validity coefficients is homogeneous.

Course length. Table 39 also indicates that the average validity coefficient is 0.26 when the course length was reported, 0.38 when the course was only one semester, and

0.28 when the course was a two semester course. The z-tests indicate that the average validity coefficient for a one semester course is significantly greater than when the course length was unknown or when the course was a two semester course.. Homogeneity tests indicate that the subset of validity coefficients for two semester courses is homogeneous ($Q_w = 7.1$ (df = 13)).

Season. Table 39 also indicates that the average validity coefficient is 0.21 when the study did not report when the evaluation was carried out, 0.45 when evaluation was carried out in the spring, and .38 when evaluation was carried out in the winter. The z-tests indicate that the average validity coefficients when studies reported when evaluation was taken are significantly different from those in studies which did not state the season in which student evaluations were collected.

Multiple regression models of course and institutional study features. Five multiple regression models were tested and are presented in Table 40. As can be seen, adding both class size and discipline (Model # 2), significantly adds to the prediction compared to only adding class size (Model # 1). Likewise, adding season (Model #3) to class size and discipline (Model # 2) significantly adds to the prediction. However, there is no significant increase when course length (Model # 4) is added. Therefore, only class size, discipline and season will be retained in the final multiple model. These three variables explain 4.9 % of the variability in validity coefficients.

Table 39. *Predictors in regression with course and institutional explanatory study features.*

FEATURE	LEVEL	k	mem	se	z test	Q_W
Type of Instruction	unknown	465	0.31	.03	12.27	845.1
	drill	32	0.38	.06	1.22	51.0
	laboratory	44	0.41	.06	2.07	171.1
	AV/tutorial	32	0.43	.09	1.52	44.4
	Disc/tutorial	39	0.19	.11	-1.22	106.9
	lecture	54	0.31	.09	-0.03	244.5
Teaching duration	unknown	498	0.32	.02	13.54	979.7
	<3 hrs	95	0.44	.06	2.40	329.3
	>3 hrs	72	0.26	.08	1.96	184.2
Discipline	arts	365	0.25	.03	9.03	778.8
	science	300	0.44	.05	6.33	693.2
Course length	unknown	230	0.26	.03	7.81	803.4
	one semester	421	0.38	.06	3.83	679.7
Season	two semesters	14	0.28	.11	0.27	7.1*
	unknown	327	0.26	.03	9.76	906.5
	spring	123	0.45	.05	4.30	148
	winter	215	0.38	.04	3.11	425
Type of institution	undergraduate	39	0.32	.07	5.01	203.4
	graduate	626	0.33	.07	0.25	1296.8
Class size	unknown	147	0.46	.05	10.33	221.8
	<400	164	0.45	.09	-0.25	298.8
	>400 and < 600	159	0.20	.08	-4.73	619.7
	> 600	195	0.31	.08	-3.08	331.4
Section size	unknown	149	0.46	.05	10.16	375.7
	< 17	170	0.36	.08	-1.87	478.6
	> 17 and < 23	152	0.23	.08	-4.24	296
	> 23	194	0.32	.08	-2.81	331.4

note k is the number of validity coefficients. MEM is the mean effect magnitude backtransformed to r. SE is the standard error about the mean Q_W. * indicates that the set is homogeneous at α=.05.

Table 40. *Multiple regression models of course and institutional explanatory study features*

#	Model Description	%exp	Q_r	df	ΔQ_r	Δdf
1	Class Size	1.9	300.8	3		
2	Class size. Discipline	3.4	322.7	4	21.9 *	1(1)
3	Class size, Discipline. Season	4.9	346.0	6	23.3 *	2(2)
4	Class size. Discipline. Season. Course length	5.1	348.3	8	2.3	2(3)
5	All predictors	6.8	373.9	15	26.1 *	8(4)

note % exp is the cumulative percent explained by the model. QE is the Goodness of Fit statistic. df are the degrees of freedom for the model $\Delta Q_r$and Δdf are the differences in the Goodness of Fit statistic and associated degrees of freedom between the indicated model and the previous model.

Multiple Regression Models

A hierarchical multiple regression model was constructed consisting of the study features retained in each of the subsets. described above.This model explained 23.8 % of the variance, but a Goodness of Fit statistic, (Q_E) of 870.7 (df = 629) indicated that there was still significant variability to be explained. This is in large part because of the large amount of missing data. For example, timing is a significant predictor; however, for almost 50% of the outcomes, the studies did not provide any information on this variable. Therefore, the variability in this subset of the data set cannot be resolved. Therefore, a subset of data was selected which excluded missing data for salient variables. That is, cases were selected only if they provided data on timing and instructor rank. Two hundred and twenty-five outcomes met this criteria.

The weighted average validity coefficient for these 225 outcomes is 0.47. The 95% confidence interval (based on the number of class sections) extends from 0.43 to 0.51. The homogeneity statistic, Q_T, was 262.6 (df = 210), indicating that the data set is heterogeneous and that study features may moderate the validity of student ratings. A hierarchical multiple regression model was constructed consisting of the following study features, entered in this sequence: structure (F1, F2, F3, F4), source of study (theses, articles), timing (after, before), group equivalence (none, statistical control, experimental control), and rank (TA, faculty). This model explained 19.3 % of the variance (r = 0.44). The Goodness of Fit statistic, (Q_E) was 210.5 (df = 200), indicating that the remaining variability was not significant. Table 41 shows the beta weights, the standard errors, and the z-tests for the predictors entered into the regression equation.

Thus, the mean validity coefficient for ratings that evaluated full faculty members' role in delivering instruction before the final exam, extracted from published studies, and with experimental control for ability differences was 0.703. The 95% confidence interval (based on the number of class sections) extends from 0.49 to 0.91. On the other hand, the mean validity coefficient for ratings that evaluated Teaching Assistant's role in delivering instruction before the final exam, extracted from theses, statistically controlling for ability differences is -.045. The 95% confidence interval (based on the number of class sections) extends from -0.267 to 0.176.

In conclusion, the published multisection validity literature suggests that under appropriate conditions (all instructors are faculty members, evaluation is carried out prior to students' knowing their final grade, sections are equivalent in terms of student ability or equivalence is experimentally controlled) and the validity coefficient is corrected for attenuation, more than 45% of the variation in student learning among sections can be explained by student perceptions of instructor effectiveness.

Table 41. *Beta-weights for optimal model (k = 225 outcomes).*

Predictors	Beta weight	SE	z-test
Intercept: Theses, After, No ability control, TA's	.180	.118	1.53
Structure: F1	.115	.047	2.44
F2	.055	.041	1.34
F3	.076	.092	.83
F4	.135	.073	1.85
Study source: Article	.430	.108	3.98
Timing: Before	-.296	.070	-4.24
Group Equivalence: Statistical control	-.044	.044	-.99
Experimental control	.205	.119	1.72
Rank: Faculty	.241	.081	2.99

note SE is the standard error. about the Beta-weight in the regression equation.

Discussion

I conducted three studies in the meta-analysis of the multsection validity studies. In the first study, I addressed the issue of whether the three first-order factors, presenting instruction, facilitating interactions, and evaluating learning, were distinct. That is, I determined whether any of the study features differentially influenced validity. Having determined that the three factors were not distinct, I subsequently coded the outcomes on the basis of the degree to which they represented *general instructional skill* (the first second-order factor) and computed the average validity. The results of this second study indicated that, in general, there is a medium correlation (.33) between student ratings and student learning. That is, instructional effectiveness explains about 10% of the variability in student learning. However, the data set is heterogeneous. Therefore, in the third study I explored the degree to which methodological and publication features, quality of evaluation features, student rating form features, achievement measure features, instructor and student explanatory features, and course and institutional explanatory features explain this heterogeneity. The results of these analyses are discussed below.

Strengths and Weaknesses of the Analyses:

Cohen (1980, 1981) extracted approximately 50% of the outcomes from the multisection validity studies. He extracted one outcome per factor per course per study. In addition, he subdivided these outcomes into ten data sets, presumably reflecting distinct rating dimensions. A consequence of these choices, is the low statistical power of the analyses and the wide confidence intervals. For example, the confidence interval about the mean for Overall Instructor extended from .21 to .61. He also used a Glassian approach and consequently could not determine the adequacy of the hierarchical regression model. Nevertheless, he was able to explain 31% of the variability with three study features: instructor rank, evaluation timing, and control for scoring bias.

Another consequence of Cohen's choice of approach is that the unit of analysis was the outcome. Each study provided one or more estimates (outcomes) of the mean population validity coefficient. However, individual studies usually vary in sample size and therefore have different sampling variances. In the Glassian approach to meta-analysis, this variation in sampling variance is not taken into account and all studies are treated equally. Thus, in the Cohen meta-analysis (Cohen, 1981) the 95% confidence interval is computed with the outcome (number of courses in the multisection validity studies) as the unit of analysis and includes the error due to the individual sampling variances.

However, I used the meta-analysis approach developed by Hedges and Olkin

(1985) in which variation in sampling variances among studies is taken into account, such that studies with smaller sample sizes (and larger sampling variances) contribute less in calculating the average. Thus, the population mean validity is a weighted average of all estimates, where each estimate is weighted by the inverse of its sampling variance. Since the 95% confidence interval does not include these individual sources of error, it is smaller than the confidence interval computed on the basis of the unweighted approach of Glass (1978). However, there is controversy on whether the number of outcomes (number of courses in multisection validity studies) or number of subjects (number of instructors in multisection validity studies) should be basis of the computation of the confidence interval (Rosenthal, 1995). Since the question I am addressing concerns the validity of student ratings of instruction across instructors and not across multisection courses, I decided to use the instructor as the unit of analysis in the computation of the 95% confidence interval. Therefore, the confidence interval is small, and the results appear to be less variable than Cohen's results indicated.

In this meta-analysis, I also decided to extract all the outcomes (741) and model the interdependencies. This avoids the possibility of introducing bias by the selection of outcomes for inclusion. With a larger data set, the statistical power of the tests is high enough to detect significant predictors. I also decided to use the homogeneity approach (Hedges, & Olkin, 1985). This approach allowed me to test not only whether study features accounted for a significant amount of the discrepancies in the findings; but also, whether the unexplained variance was significant.

Feldman (1989) used a univariate approach in his meta-analysis of the multisection validity studies. However, he replicated validity coefficients that come from factor scores that include more than one instructional category in order to compute the mean validity for each instructional category. Although, he corrected for this inflation in total number of entries by weighting each outcome by the inverse of the number of entries, he did not take the interdependencies into account. Thus, it is not possible to state whether the values for any two categories are significantly different. Moreover, defining different instructional categories and calling them dimensions, does not insure that they are empirically distinct. Feldman neither determined whether the data sets were homogeneous; nor investigated the influence of study features on the validity of these instructional categories. In the studies reported in this thesis, I modelled, the interdependencies, used a multivariate homogeneous approach and investigated the influence of study features on the validity of general instructional skill. It is therefore, difficult and meaningless to compare my results to Feldman's results. Not only, did I use different approaches, I also asked different questions. Moreover, Feldman did not investigate the influence of study features on the validity of student ratings.

There are a number of limitations with the analyses reported in this thesis. Firstly, there is a large degree of dependency within the data set. Forty-three studies yielded 741 validity coefficients that were distributed in 88 independent samples. I modelled these dependencies by estimating the covariance matrix using the internal consistency (.78). of the first second order factor. However, in the absence of Monte Carlo studies, I do not have a measure of the adequacy of this modelling. Secondly, there is a high degree of collinearity among the study features. For example, most faculty taught courses in which there was some coordination or they had complete autonomy; while most TA's taught courses which were directed. Thus, it is impossible to unambiguously interpret the results of the regression analysis. Thirdly, a meta-analysis is a correlational study in which there is no experimental control. Many of the study features do not have an equal (or even approximately equal) number of cases for each level of the variable. In many cases, the level with the largest number of cases is the level representing missing data. Despite these limitations, present in most meta-analyses, the analyses reported here confirm the results reported by Cohen (1981), and in addition, demonstrate that all the significant variability in data from studies reporting the values of key study features, can be explained by a small set of study features.

Study 1: Presence of Significant and Practically Important Interactions

The factor analysis of the aggregated student rating forms (or rather the reproduced correlation matrices) indicates that students judge instructional effectiveness on the basis of general instructional skill (the second-order factor) that consists of three correlated first-order factors (presenting material, facilitating interactions, and evaluating learning) and a miscellaneous factor. In order to investigate the possible distinctiveness of the three first-order factors comprising general instructional skill, I searched for significant and practically important interactions between the first-order structure and 38 study features. I reasoned that if these factors were distinct, many of the study features would have a differential influence on the three factors. For example, I might expect that timing of evaluation would influence the validity of students' evaluation of presenting material, facilitating interactions, and evaluating learning differentially. Students can rate their instructors' presentation after one class, however; it takes time for the instructor to facilitate interactions in the classroom, and students can only rate their instructors' evaluation of learning after such evaluation has been given. However, of the 38 possible interactions, only 4 were both significant and explained more than 1% of the variability in validity. These were study source, type of instruction, instructor autonomy, and student gender. Thus, it appears that the validity of student ratings of instruction is generalizable across most courses, instructors, and students used in multisection validity studies.

As discussed in Study 3, the data extracted from theses appears to often contain errors, and therefore the high negative (-.53) validity coefficient for student ratings of the instructors' ability to facilitate interactions may reflect the low reliability of the extracted outcomes. Therefore, the first interaction will not be discussed further.

There is a significant and practically important interaction between type of instruction and the first-order factor structure. Figure 2 shows that the validity of student ratings is significantly lower when students in lectures rate their instructors' ability to facilitate interactions compared to students in classes stressing drill, laboratory sections, or discussion groups. This may arise because there may be very little interaction in lecture courses. Thus, there may be too few high scores for facilitating interactions in lecture courses. This range restriction would attenuate the association between student ratings and student learning. Courses stressing drill (and practice), laboratory, and discussion courses, on the other hand permit a much greater range of interactions. Thus, students' ratings may also exhibit a better range of scores. An alternative explanation is that the interaction may indicate that facilitating interactions does not contribute to student learning in lecture classes to the extent that it does in other classes; while, presenting material and evaluating learning influence student learning in all types of courses.

There is also a significant and practically important interaction between instructional autonomy and the first order factor structure. Figure 3 shows that the validity of student ratings is significantly lower when students rate instructors with complete autonomy on their ability to facilitate interaction compared to that when students rate instructors with no or little autonomy, or when students rate instructors on presenting material and evaluating learning. Figure 3 shows that instructors in multisection courses where there is complete autonomy (no interactions among instructors) may have been selected to teach these courses, instead of being selected to teach multisection courses requiring co-ordination, because they do not interact well. Thus, there would be a reduced range in student ratings of facilitating interactions resulting in an attenuation of the validity coefficient. Alternatively, multisection courses in which instructors are completely autonomous may be restricted to specific courses (*e.g.*, higher-level, graduate courses) in which facilitating interactions is not as important.

There is also a significant and practically important interaction between student gender and the first order factor structure. Figure 4 shows that the validity of student ratings is significantly lower when male students rate their instructors' ability to facilitate interactions compared to when they rate their instructors' ability to either present material or evaluate learning. On the other hand, it is significantly lower when female students rate their instructors' ability to present material compared to when they rate their instructors' ability to either facilitate interactions or evaluate learning. Male students may compensate for their instructors' "deficiencies" in presenting material and therefore learn the course material. The association between instructional effectiveness (presenting material) and learning would be attenuated for male students. On the other hand, female

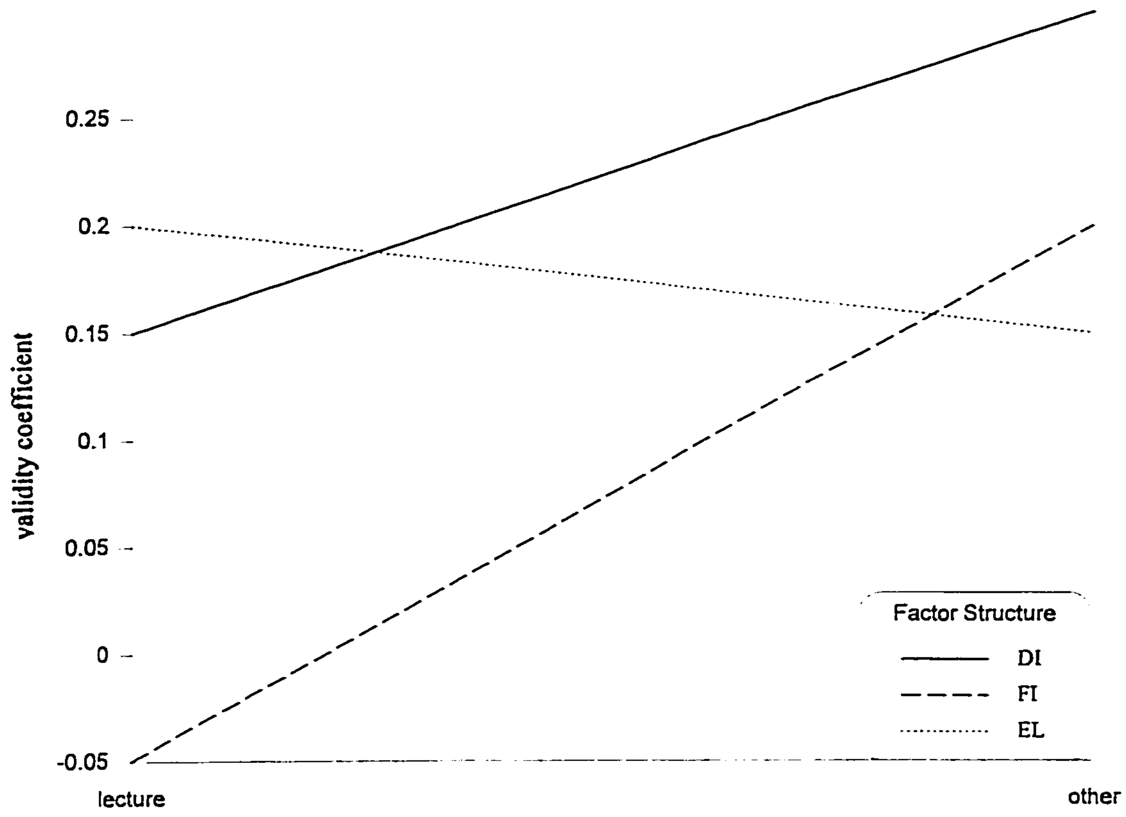


Figure 2. Interactions between First-order Factor Structure and Type of Instruction

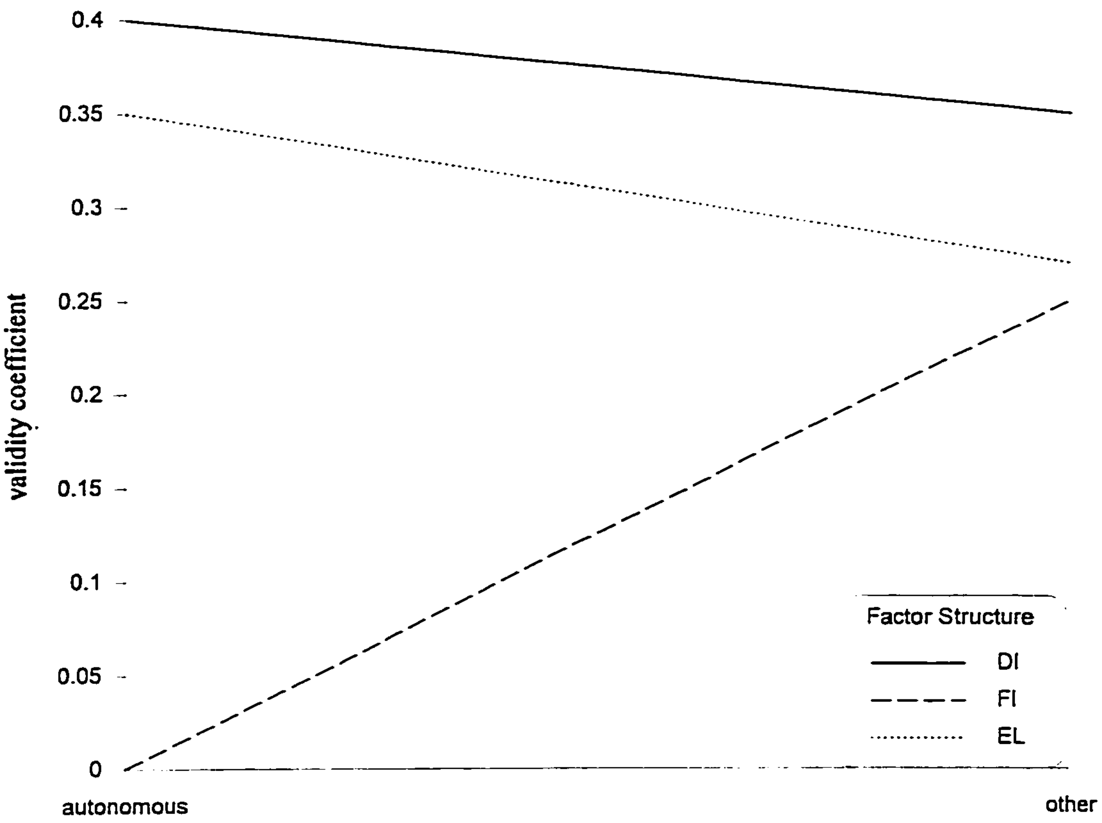


Figure 3. Interactions between First-order Factor Structure and Instructional Autonomy

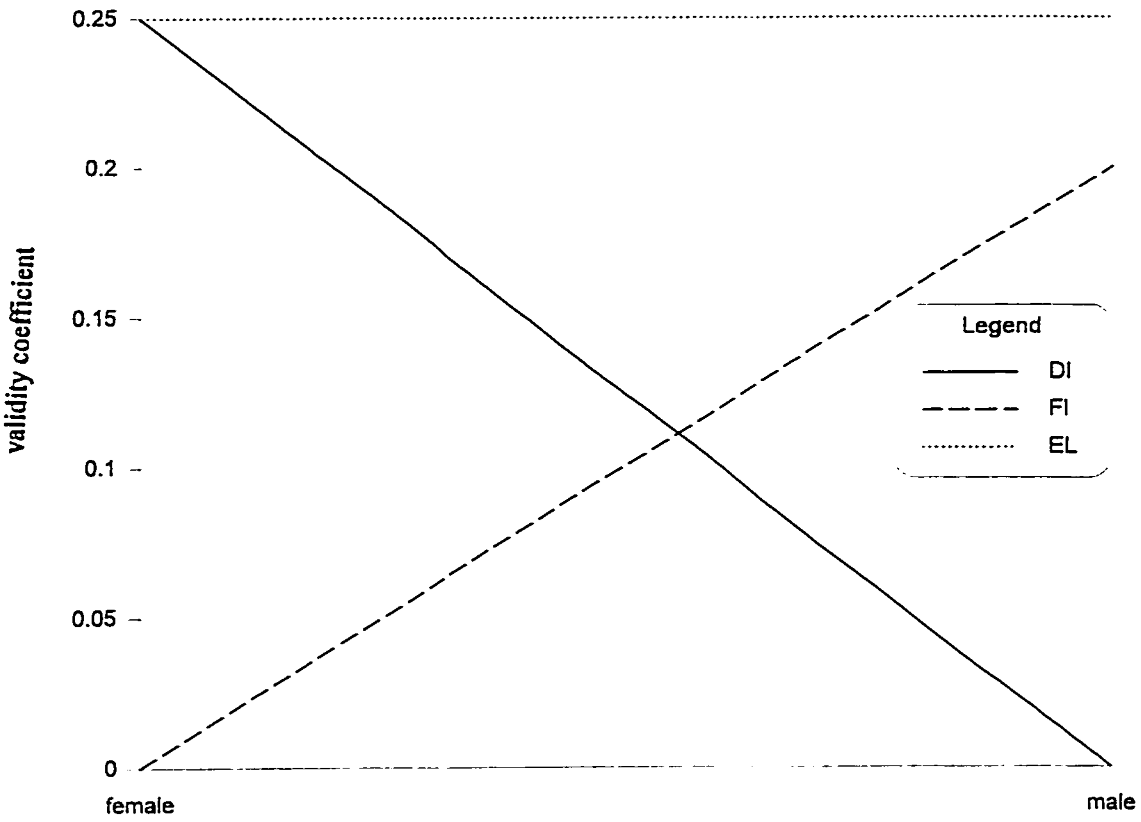


Figure 4. Interactions between First-order Factor Structure and Student Sex

students may compensate for their instructors' "deficiencies" in facilitating interactions and learn the course material. Thus the association between instructional effectiveness (facilitating interactions) and learning would be attenuated for female students. On the other hand, there may be no association between an instructor's ability to present material well and learning for male students; and no association between an instructor's ability to facilitate interactions and learning for female students. This may mean that when instructors work at facilitating interactions, at the expense of presenting material clearly, they enhance learning for male students but not female students and *vice versa*. Thus, to reduce differential effects, instructors should attend to all aspects of effective instruction.

Study 2: Overall Validity of Student Ratings of Instruction

Overall research supports the validity of student ratings of instruction. When attenuation due to the unreliability of the instruments is taken into account, at least 25% of the variability in student learning across sections of a multisection course can be explained by differences in the general instructional skill of the instructors. Thus, in general, differences in ratings of instruction reflect differences in instructor mediated student learning.

However, the multisection validity studies that comprise this data set were conducted primarily in large introductory courses. The conclusion may or may not generalize to smaller, graduate, or other types of courses. Similarly, the criterion measure, in most cases, was a common final multiple choice examination. The conclusions may not generalize to different measures of student learning. Moreover, the data set is heterogeneous; therefore, the validity of student ratings is moderated by extraneous factors. These will be discussed below.

Study 3: Influence of Study Features

The results of this analysis indicate that 23 of 39 study features significantly moderate the validity of student ratings of instruction. These study features, individually, explain from about 0.5% to 9% of the variability in validity. However, there is a large degree of collinearity among the data. Thus, including all the significant predictors into a multiple regression model explains only about 30% of the variability in the complete data set. This is, in large part, due to the large amount of information about study features that is missing from the data set.

Methodological and Publication Features

The way in which the primary researchers computed validity coefficients and the source of the study account for approximately 5% of the variability in reported outcomes. Studies employing rank correlations as measures of validity and/or studies reported in theses are significantly different from other studies. There were only 19 outcomes computed for rank correlations in the data set. The primary researchers used rank rather than Pearson product correlation because their sample sizes were too small. Thus, the mean validity coefficient (.54) from this subset of the data may be unreliable and should not be included in estimating the overall validity of student ratings. Similarly, the mean validity coefficient calculated for outcomes from theses (-.28) should also be discounted.

Many of the theses contained typographical, statistical. and other errors.

Quality of Evaluation Study Features

Study features which reflect the quality of the evaluation procedure, such as whether the evaluation was carried out before or after the end of the semester, whether instructors graded their own students' exams (used as the criterion measure), whether instructors had prior knowledge of this test, and whether section equivalences in ability were controlled, accounted for approximately 6% of the variability in reported outcomes.

In agreement with Cohen (1980. 1981, 1982, 1983), I found that the validity coefficients from studies in which instructional evaluation is carried out during or after the last week of the semester are significantly higher than those from studies carried out earlier in the semester. Students may be "rewarding" instructors who have given them high grades, or they may be using grades, rather than instructional behaviours, as measures of instructional effectiveness. In both cases, the relationship between instructional behaviour and student learning is confounded and consequently the validity of student ratings of instruction cannot be assessed unambiguously. Therefore, to preclude the possibility the students are rewarding lenient professors, the evaluation of instructional effectiveness should be conducted in the middle. rather than at the end of the term.

In agreement with Cohen (1980. 1981, 1982, 1983), I also found that validity coefficients from studies in which researchers, departmental committees, or external evaluators graded student exams are significantly higher than when instructors graded their own students' exams. Instructors may temper their assessment of student performance for both appropriate and inappropriate reasons. In either case, this non-uniform grading practice will attenuate the validity coefficient. Similarly, instructors who know what will be on the common achievement test may inadvertently pass this information on to students, who may subsequently outperform students in other sections. Thus, student performance will not accurately reflect instructor effectiveness but rather prior knowledge of test items.

Section to section differences in performance may also reflect differences in initial student ability rather than differences in instructor effectiveness. The results also indicate that the studies in which initial differences do not exist or in which initial differences are experimentally controlled by random assignment report *consistent*, high, validity coefficients (in the order of .50). In conclusion, the multsection validity studies indicate that when student ratings are conducted under rigorous conditions, controlling for the quality of the evaluation procedure, student ratings of general instructional skill are valid.

Student Rating Form Study Features

Characteristics of the rating form account for about 14% of the variability in reported validity coefficients. The most important predictor is the dimensional structure of the factor score. That is, the distribution of the items, defined in terms of the four first-order factors, contributing to the computation of the validity coefficient moderates the size of the validity coefficient. The inclusion of items pertaining to *Disciplinary Actions*, *Choice of Required Materials*, *Knowledge of Domain*, and *Use of Objectives* attenuates the validity coefficient. The influence of structure is also reflected in the influence of such study features as the diversity index, factor length, and form length. That is, studies based on short forms, factor scores based on fewer items, and unidimensional scales yield higher validity coefficients than do long forms, factor scores based on a large number of items, or multidimensional scales. Since global ratings have higher validity coefficients, and since these items are more likely to be absent from long forms using multidimensional scales, the negative relationships my reflect the collinearity in the data set, rather than the true influence on number of items, *etc.*, on the validity of student rating forms.

Achievement Measure Study Features

Characteristics of the achievement measure appear to account for only 3 % of the variability in reported outcomes. Very few of the characteristics appear to be significant predictors of validity. This may reflect a problem with *range restriction* in the achievement measure. Study 1 also indicated that there were more interactions between study features and the first-order structure than with the other subsets of study features. Thus, interactions with the first order factor structure may obscure the influence of achievement measure study features and general instructional skill. However, most of the primary studies either do not report on characteristics of the achievement measure, or use only one level. Thus, the very little is known about the influence of characteristics of the achievement test on student ratings of instruction..

Instructor and Student Explanatory Study Features

Characteristics of the instructor, such as rank, experience, and autonomy, account for about 5% of the variability in reported outcomes. Rank, experience and autonomy are significant predictors, but these three characteristics are highly correlated. In this data set, most faculty are experienced and have moderate to full autonomy; while most TA's are inexperienced and have no or moderate autonomy. Cohen (1981, 1982, 1983) also found that these instructor variables were important moderators of the validity of student ratings. The validity of student ratings are significantly higher for experienced faculty teaching courses that are co-ordinated than they are for inexperienced TA's teaching similar courses. Hativa and Raviv (1993) also found that global ratings accurately reflect the different dimensions of instruction for full faculty but not for teaching assistants. Thus, caution should be used in comparing the results of student ratings across faculty differing in rank or experience.

Course and Institutional Explanatory Study Features

Characteristics of the course account for about 5% of the variability in reported outcomes. A number of researchers (Centra, & Creech, 1976; Neumann, & Neumann, 1985) reported that student evaluations were higher in the soft (arts) as opposed to the hard (science and math) disciplines. However, studies conducted in science and math classes report significantly higher validity coefficients. Perhaps grades are lower in the soft disciplines and therefore, the relationship between student ratings and student achievement is attenuated. However, an alternative explanations for this finding may be that discipline and some of the other characteristics are collinear. For example, more science and math courses may have full faculty as instructors compared to arts courses. Science and math courses often employ skill or problem-solving achievement tests, *etc.*

Feldman (1978) and Marsh (1987) suggested that evaluations are rarely compared across disciplines, and that therefore, discipline differences do not matter in practice. However, multidisciplinary courses such as research methodology are often taught by faculty from different disciplines (mathematics, psychology, sociology, economics, *etc.*). In such courses, evaluations would be compared across disciplines. Thus, care must be exercised in such comparisons.

Multiple regression Model

One of the limitations of a meta-analysis is that the reviewer cannot collect more data. Many of the primary researchers did not report characteristics of their study that have subsequently been shown to be important. For example, many studies did not report the rank of the instructor (47%) or when the evaluation was carried out (45%). Thus an important source of variability cannot be assessed in these studies. To get an accurate measure of the degree to which knowledge of study features explains the variability in study findings, studies with missing data must be excluded from the analysis. Thus, I selected a subset of the data which provided complete data on these two study features. The mean validity coefficient for this subset was moderately high (.47). Moreover, five study features (dimensional structure, source of study, timing, control for group equivalence, and rank) explained the significant variability. Any unexplained variability could be accounted for by sampling error. Thus, one can conclude that student ratings of instruction are *consistently* high in well controlled studies on full faculty. The validity of student ratings of instruction reported in published articles for full faculty conducted at the beginning of the semester in classes controlled for group equivalence is .70. That is, in this subset of the data, almost 50% of the variability in student performance between sections can be predicted on the basis of instructor effectiveness.

General Conclusions and Recommendations for Future Research

Student ratings of instruction are widely used in post-secondary institutions to assess the effectiveness of instruction. They are used to aid students in course selection, to provide instructors with feedback for course and instructional improvement, to provide researchers with information on the teaching-learning process, and to provide administrators with information for hiring and tenure decisions. In general, student ratings have been shown to be reliable, valid, and useful measures of instructor effectiveness. Although individual student rating forms can include many specific instructional behaviours, students, across rating forms, assess general instructional skill. General instructional skill is a composite of three correlated factors, delivering instruction, facilitating interactions, and evaluating learning.

The research summarized in this paper supports the view that global ratings, or a single score representing general instructional skill should be used for summative evaluations. Single scores representing general instructional skill can be generated by weighting items by their loadings on either the principal components matrix (Table 9) or factor structure matrix (Table 10). However, I would not recommend this method of generating a weighted score. First, not all instructional dimensions (and all items) are equally valid. Second, the instructional dimensions that were shown to be valid in the multisection validity studies are not necessarily valid in other courses (*e.g.*, small higher level courses). Since the global instructor dimension has a high loading (.94) on the first principal component, I would recommend that summative evaluations be based on a number of items assessing overall instructor effectiveness.

I would also recommend that three composite scores be generated, reflecting the three instructional roles; delivering instruction, facilitating interactions, and evaluating learning. These scores can be generated by weighting the scores for individual items on a student rating form according to the factor structure matrix (see Table 10). For example, if an instructor received scores of 3.5, 2.5, and 4 on items assessing evaluation, feedback, and a friendly classroom environment, the scores could be weighted by .4, .4, and .2 respectively to generate a composite score of 3.2 reflecting the instructor's role in evaluating learning. These specific scores could be used to identify teaching strengths and weakness and be useful to the instructor for instructional improvement.

When student ratings are to be used for summative decisions, it is important that the circumstances under which student ratings are collected and analyzed be appropriate and rigorously applied. Centra (1993) recommended that when student ratings are used for purposes of summative evaluation, students should anonymously complete short rating forms in class at the end of the semester but before final grades are known. The instructors should neither be in the classroom, nor collect the student rating forms. The research presented in this thesis indicates that the structure and length of the student rating form, and the timing of the evaluation influences the validity of the student rating of general instructor skill. However, other administrative conditions, such as student anonymity, the stated purpose of evaluation, or whether the instructors themselves carried out the evaluation procedure, did not significantly influence the validity of student ratings of instruction. Nevertheless, standardized procedures should be used for ethical, legal, and practical reasons.

The consensus has been that biasing variables play a minor role in student ratings of instruction (Marsh, 1987; Murray, 1984). However, the research summarized in this paper indicates that instructor and course variables can influence the validity of student ratings of instruction. For example, students rate instructors more accurately when they evaluate full-time faculty teaching large science courses compared to when they evaluate teaching assistants teaching medium classes in a language course. Since the instructor has no control over these biasing factors, ratings should either be statistically controlled for these variables or an instructor's rating should only be compared to a norm group with the same characteristics. However, there are a number of problems with the use of norm groups (Abrami, 1993; Hativa, 1993; McKeachie, 1996). For example, in some cases the norm group would be so small that the data are unreliable. Abrami and McKeachie have also argued that such norms have negative effects on faculty morale. By definition, half the faculty would be below the norm, yet they could be excellent teachers. Norms also rank instructors inappropriately because individual differences in rank may have no practical significance. Although the above factors appear to bias student ratings, the multisection validity studies do not (and cannot) indicate whether the influence of the factor is on student ratings or on student learning, as measured by a common achievement test.

Faculty often express concerns that administrators interpret student ratings of instruction without taking contextual factors into consideration. Thus they believe that decisions made on the basis of student ratings are unfair. Moreover, faculty are increasingly experimenting with different pedagogical methods (*e.g.*, small and large class lecturing, tutoring and advising, studio classes, discussion and small group methods including cooperative learning, individualized and mastery learning,). Student ratings, designed with the lecture format in mind, may not be appropriate for these instructional contexts. Instructors will only trust evaluations of their performance based on student ratings when contextual factors, including instructional style, are considered. Because of the limitations of the research on contextual factors, student ratings should not be "overinterpreted" . That is, only crude judgements of instructional effectiveness (exceptional. adequate, and unacceptable) should be made on the basis of student ratings.

References

Abrami, P.C. (1984, February). Using meta-analytic techniques to review the instructional evaluation literature. *Postsecondary Education Newsletter*, **6**, 8.

Abrami, P. (1985). Dimensions of effective college instruction. *Review of Higher Education*, **8**, 211-28.

Abrami, P. C. (1988). *The dimensions of effective college instruction.* Montreal.: Concordia University.

Abrami, P. C. (1989). Seeking the truth about student ratings of instruction. *Educational Researcher*, , 43-45.

Abrami, P.C., Cohen, P.A., & d'Apollonia, S. (1988). Implementation problems in meta-analysis. *Review of Educational Research*, **58**, 151-179.

Abrami, P.C., & d'Apollonia, S. (1990). The dimensionality of ratings and their use in personnel decisions. In M. Theall, & J. Franklin (Eds.) *Student ratings of instruction: Issues for improving practice. New directions for teaching and learning.* Number 43, pp. 97-111. San Francisco: Jossey-Bass.

Abrami, P. C., & d'Apollonia, S. (1991). Multidimensional students' evaluations of teaching effectiveness- generalizability of "N=1" research: Comment on Marsh (1991). *Journal of Educational Psychology*, **83**, 411-415.

Abrami, P.C., d'Apollonia, S., & Cohen, P.A. (1990). The validity of student ratings of instruction: What we know and what we do not. *Journal of Educational Psychology,* **82**, 219-231.

Abrami, P.C., d'Apollonia, S., & Rosenfield, S. (1996). The dimensionality of student

References

Abrami, P.C. (1984, February). Using meta-analytic techniques to review the instructional evaluation literature. *Postsecondary Education Newsletter*, **6**, 8.

Abrami, P. (1985). Dimensions of effective college instruction. *Review of Higher Education*, **8**, 211-28.

Abrami, P. C. (1988). *The dimensions of effective college instruction*. Montreal.: Concordia University.

Abrami, P. C. (1989). Seeking the truth about student ratings of instruction. *Educational Researcher*, , 43-45.

Abrami, P.C., Cohen, P.A., & d'Apollonia, S. (1988). Implementation problems in meta-analysis. *Review of Educational Research*, **58**, 151-179.

Abrami, P.C., & d'Apollonia, S. (1990). The dimensionality of ratings and their use in personnel decisions. In M. Theall, & J. Franklin (Eds.) *Student ratings of instruction: Issues for improving practice. New directions for teaching and learning*. Number 43, pp. 97-111. San Francisco: Jossey-Bass.

Abrami, P. C., & d'Apollonia, S. (1991). Multidimensional students' evaluations of teaching effectiveness- generalizability of "N=1" research: Comment on Marsh (1991). *Journal of Educational Psychology*, **83**, 411-415.

Abrami, P.C., d'Apollonia, S., & Cohen, P.A. (1990). The validity of student ratings of instruction: What we know and what we do not. *Journal of Educational Psychology*, **82**, 219-231.

Abrami, P.C., d'Apollonia, S., & Rosenfield, S. (1996). The dimensionality of student

References

Abrami, P.C. (1984, February). Using meta-analytic techniques to review the instructional evaluation literature. *Postsecondary Education Newsletter*, **6**, 8.

Abrami, P. (1985). Dimensions of effective college instruction. *Review of Higher Education*, **8**, 211-28.

Abrami, P. C. (1988). *The dimensions of effective college instruction.* Montreal.: Concordia University.

Abrami, P. C. (1989). Seeking the truth about student ratings of instruction. *Educational Researcher*, , 43-45.

Abrami, P.C., Cohen, P.A., & d'Apollonia, S. (1988). Implementation problems in meta-analysis. *Review of Educational Research*, **58**, 151-179.

Abrami, P.C., & d'Apollonia, S. (1990). The dimensionality of ratings and their use in personnel decisions. In M. Theall, & J. Franklin (Eds.) *Student ratings of instruction: Issues for improving practice. New directions for teaching and learning.* Number 43, pp. 97-111. San Francisco: Jossey-Bass.

Abrami, P. C., & d'Apollonia, S. (1991). Multidimensional students' evaluations of teaching effectiveness- generalizability of "N=1" research: Comment on Marsh (1991). *Journal of Educational Psychology*, **83**, 411-415.

Abrami, P.C., d'Apollonia, S., & Cohen, P.A. (1990). The validity of student ratings of instruction: What we know and what we do not. *Journal of Educational Psychology*, **82**, 219-231.

Abrami, P.C., d'Apollonia, S., & Rosenfield, S. (1996). The dimensionality of student

ratings of instruction: What we know and what we do not. Invited article for *Higher Education: Handbook of Theory and Research.* Vol. **11**, 213-263.

Abrami, P.C., Dickens, W.J., Perry, R.P., & Levinthal, L. (1980). Do teacher standards for assigning grades affect student evaluations of instruction? *Journal of Educational Psychology*, **72**, 107-118.

Abrami, P.C., Levinthal, L., & Perry, R.P. (1982). Educational seduction. *Review of Educational Research*, **52**, 446-464.

Abrami, P.C., & Mizener, D.A. (1985). Student/instructor attitude similarity, student ratings, and course performance. *Journal of Educational Psychology*, 77, 693-702.

Aleamoni, L.M. (1981). Students ratings of instruction. In J. Millman (Ed.), *Handbook of teacher evaluation*. Beverly Hills, CA: Sage.

Aleamoni,L.M.(1987). Techniques for evaluating and improving instruction. *New Directions for Teaching and Learning*. **31**. Jossey-Bass. San Francisco.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *The standards for educational and psychological testing.* Washington,DC: American Psychological Association.

Anderson, J.R. (1983). The architecture of cognition. Cambridge, MA: Harvard University Press.

Anderson, H.H. & Jacobson, A. (1965). Effect of stimulus inconsistency and discounting instruction in personality impression formation. *Journal of Personality and Social Psychology*, **4**, 531-539.

Balzer, W.K., & Sulsky, L.M. (1992). Halo and performance: A critical examination. *Journal of Applied Psychology*, **77**, 975-985.

Bankgert-Drowns, R.L. (1986). A review of developments in meta-analytic methods. *Psychological Bulletin*, **99**, 388-v399.

Becker, B.J. (1992). Models of science achievement: Factors affecting male and female performance in school science. L.V. Hedges, R.J. Light, T.A. Louis, & F. Mosteller, *Meta-analysis for explanation: A casebook.* New York: Russell Sage Foundation.

Becker, B.J., & Schram, C.M. (1994). Examining explanatory models rhrough research synthesis. In H. Cooper, & L.V. Hedges (Eds)*The handbook of research synthesis.* New York: Russell Sage Foundation.

Becker, B.E., & Cardy, R.L. (1986). Influence of halo error on appraisal effectiveness: A conceptual and empirical reconsideration. *Journal of Applied Psychology*, **71**, 662-671.

Bendig, A.W. (1953a). The relation of level of course achievement to students' instructor course ratings in introductory psychology. *Educational and Psychological Measurement,* **13**, 437-448.

Bendig, A.W. (1953b). Student achievement in introductory psychology and student ratings of the competence and empathy of their instructors. *Journal of Psychology,* **36**, 427-433.

Bendig, A.W. (1954).A factor analysis of student ratings of psychology instructors on the Purdue Scale. *Journal of Educational Psychology*, **45**, 385-393.

Benton, S.E., & Scott, O. (1976, April). A comparison of the criterion validity of two types of student response inventories for appraising instruction. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco. (ED # 128-397).

Bolton,A. (1990). Personal communications.

Bolton, A., Bonge, D., and Marr, J., (1979). Ratings of instruction, examination performance, and subsequent enrollment in psychology courses. *Teaching of Psychology*, **6**, 82-85.

Borg, W.R., & Hamilton, E.R. (1956). Comparison between a performance test and criteria of instructor effectiveness. *Psychological Reports*, **2**, 111-116.

Borman, W. (1974). The rating of individuals in organizations: An alternative approach. *Organizational Behavior and Human Performance*, **12**, 105-124.

Borich, G. D. (1977). *The appraisal of teaching: Concepts and processes*. Don Mills, ON: Addison-Wesley.

Braskamp, L.A., Caulley, D., & Costin, F. (1979). Student ratings and instructor self-ratings and their relationship to student achievement. *American Educational Research Journal*, **16**, 295-306.

Bryant, F.B., & Wortman, P. M. (1984). Methodological issues in the meta-analysis of quasi-experiments. *New Directions for Program Evaluation*, **24**, 5-24.

Bryson, R. (1974). Taecher evaluations and student learning : A reexamination. *The Journal of Educational Research*, **68**, 12-14.

Bushman, B. J., Cooper, H. M., & Lemke, K. M. (1991). Meta analysis of factor analyses: An illustration using the Buss Durkee Hostility Inventory Special Issue:

Meta analysis in personality and social psychology. *Personality and Social Psychology Bulletin*, **6**, 344-349.

Cadwell, J., & Jenkins, J. (1985). Effects of the semantic similarity of items on student ratings of instruction. *Journal of Educational Psychology*, **77**, 383-393.

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

Campbell, D.T., & Stanley, J.C. (1966). Experimental and quasi- experimental designs for research: Houghton-Mifflin.

Carlberg, C., & Kavale, K. (1980). The efficacy of special versus regular class placenent for exceptional children: A meta-analysis. *The Journal of Special Education*, **14**, 295-309.

Cashin, W. E., & Downey, R. G. (1992). Using global student rating items for summative evaluation. *Journal of Educational Psychology*, **84**, 563-572.

Cashin, W. E., Downey, R. G., & Sixbury, G.R. (1994). Global and specific ratings of teaching effectiveness and their relation to course objectives: A reply to Marsh (1994). *Journal of Educational Psychology*, **86**, 649-657.

Center, B. A., Skiba, R. J., & Casey, A.(1986). A methodology for the quantitative synthesis of intra-subject design research.*The Journal of Special Education*. **19**, 387-400.

Centra, J. (1977). Student ratings of instruction and their relationship to student learning. *American Educational Research Journal*, **14**, 17-24.

Centra, J. (1979).*Determining faculty effectiveness*. San Francisco: Jossey-Bass.

Centra, J. A., & Creech, F.R. (1976). *The relationship between student, teacher, and course characteristics and student ratings of teacher effectiveness.* (Project Report 76-1). Princeton, NJ: Educational Testing Service.

Chase, C.I., & Keene, J.M. (1979). *Validity of student ratings of faculty.* Bloomingdale, Indiana: Bureau of Educational Studies and Testing, Indiana University. (ED# 169 870).

Chau, H. (1994, April). Higher-order factor analysis of multidimensional students' evaluations of teaching effectiveness. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. (ED372110)

Cochran, W. G. (1937). Problems arising in the analysis of a series of similar experiments. *Journal of the Royal Statistical Society* (Supplement), **4**, 102-118.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences.* Revised edition. Academic Press: New York.

Cohen, P.A. (1980). *A meta-analysis of the relationship between student ratings of instruction and student achievement.* Unpublished doctoral dissertation. The University of Michigan, Ann Arbor.

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

Cohen, P.A. (1982). Validity of student ratings in psychology courses: A meta-analysis of multisection validity studies. *Teaching of Psychology*, **9**, 78-82.

Cohen, P. A. (1983). Comment on a selective review of the validity of student ratings of teaching. *Journal of Higher Education*, **54**,448-458.

Cohen, P.A. (1986, April). *An updated and expanded meta-analysis of multisection student rating validity studies*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco. CA.

Cohen, P.A. (1987, April). *A critical analysis and reanalysis of the multisection validity meta-analysis*. Paper presented at the annual meeting of theEducational Research Association, Washington, D.C.

Cohen, S.H., & Berger, W.G. (1970, August). Dimensions of students' ratings of college instructors underlying subsequent achievement on course examinations. *Proceedings of the 78th Annual Convention of the American Psychological Association*, **5**, 605-606.

Cook, T.D., & Campbell, D.T. (1979). *Quasi-experimentation: Design, & analysis issues for field settings.* Boston: Houghton Mifflin.

Cook, T. D., Cooper, H., Codray, D. S., Hartmann, H., Hedges, L. V., Light, R. J., Louis, T. A., & Mosteller, F. (1992). *Meta- analysis for explanation*. New York: Sage.

Cook T., & Leviton L.(1980). Reviewing the literature: A comparison of traditional methods with meta-analysis. *Journal of Personality*, **48**, 449-472.

Cooper H.M.(1981). On the significance of effects and the effects of significance. *Journal of Personality and Social Psychology*, **41**, 1013-1018.

Cooper H.M. (1982). Scientific guidlines for conducting research integrative research reviews. *Review of Educational Research*, **52**, 291-302.

Cooper, H.M. (1979). Statistically combining independent studies: Meta-analysis of sex differences in conformity research. *Journal of Personality and Social Psychology*, **37**, 131-146.

Cooper, H.M. (1984). *The integrative research review*. Beverly Hills, CA:Sage.

Cooper, H.M., & Hedges, L.V. (1994). Research synthesis as a scientific enterprise. In H. Cooper, & L.V. Hedges (Eds)*The handbook of research synthesis.* New York: Russell Sage Foundation.

Cooper, W.H. (1981). Ubiquitous halo. *Psychological Bulletin*, **90**, 218-244.

Costin, F. (1978). Do student ratings of college teachers predict student achievement ? *Teaching of Psychology*, **5**, 86-88.

Costin, F., Greenough, W.T., & Menges, R.J. (1971). Student ratings of college teaching: Reliability, validity, and usefulness. *Review of Educational Research*, **41**, 511-536.

Cranton, P., & Smith, R. (1990). Reconsidering the unit of analysis: A model of student ratings of instruction. *Journal of Educational Psychology*, **82**, 207-212.

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, **52**, 281-302.

Crooks, T.J., & Smock, H.R. (1974). *Student ratings of instructors related to student achievement*. Urbana. Ill.: Office of Instructional Resources, University of Illinois.

d'Apollonia, S., & Abrami, P.C. (1987, April). *An empirical critique of meta-analysis: The literature on student ratings of instruction.* Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C.

d'Apollonia, S., & Abrami, P.C.(1988, April). *The literature on student ratings of instruction: Yet another meta-analysis.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

d'Apollonia, S., Abrami, P., & Rosenfield, S. (1993, April). *The Dimensionality of Student Ratings of Instruction: A Meta-Analysis of the Factor Studies.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Dowell, D.A., & Neal, J.A. (1982). A selective review of the validity of student ratings of teaching. *Journal of Higher Education*, **53**, 51-62.

Doyle, K.O., Jr. (1981). Validity and perplexity: An incomplete list of disturbing issues. *Instructional Evaluation*, **6**, 23-25.

Doyle, K. O., & Crichton, L.I. (1978). Student, peer, and self-evaluations of college instructors. *Journal of Educational Psychology*, **5**, 815-826.

Doyle, K.O., & Whitely, S.E. (1974). Student ratings as criteria for effective teaching.*American Educational Research Journal*, **11**, 259-274.

Elliott, D.N., (1950). Characteristics and relationships of various criteria of college and university teaching. *Purdue University Studies in Higher Education*, **70**, 5-61.

Ellis. N.R., and Rickard, H.C., (1977). Evaluating the teaching of introductory psychology. *Teaching of Psychology*, **4**, 128- 132.

Endo, G.T., & Della-Piana, G. (1976) A validation study of course evaluation ratings. *Improving College and University Teaching*, **24**, 84-86.

Eysenck H. J.(1978). An exercise in mega-silliness. *American Psychologist,* **33,** 57.

Eysenck H. J.(1984). Meta-analysis: An abuse of research integration. *Journal of Special Education*, **18**, 41-59.

Feldman, K. A.,.(1971). Using the work of others: Some observations on reviewing and integrating. *Sociology of Education*, **4**, 86-102.

Feldman, K.A. (1976). The superior college teacher from the student's view. *Research in Higher Education*, **5**, 243-288.

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

Feldman, K.A. (1978). Course characteristics and college students' ratings of their teachers and courses. What we know and what we don't. *Research in Higher Education*, **9**, 199-242.

Feldman, K.A. (1983). Seniority and experience of college teachers as related to evaluations they receive from students. *Research in Higher Education*, **18**, 3-214.

Feldman, K.A. (1984). Class size and college students' evaluations of teachers and courses: A closer look. *Research in Higher Education*, **21**, 45-116.

Feldman, K.A. (1986). The perceived instructional effectiveness of college teachers as related to their personality and attitudinal characteristics. *Research in Higher Education*, **24**, 139-213.

Feldman, K.A. (1988). Effective college teaching from the students' and faculty's view: Matched or mismatched priorities? *Research in Higher Education*, **28**, 291-344.

Feldman, K.A. (1989). The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education*, **30**, 583-645.

Feldman, K.A. (1990). An afterword for " The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies". *Research in Higher Education*, **31**, 315-318.

Feldman, K. A. (in press). Reflections on the study of effective college teaching and student ratings: One continuing quest and two unresolved issues. In J. C. Smard (Ed.). Higher education: Handbook of theory and research.

Fisher, R.A. (1932). *Statistical methods for research workers* (4th ed.). London: Oliver, & Boyd.

Fiske, S.T., & Neuberg, S.L.(1990). A continuum of impression formation, from category-based to individuating process: Influences of information and motivation on attention and interpretation. *Advances in Experimental Social Psychology*, **23**, 1-74.

Franklin, J.L., & Theall, M. (1990). Communicating Student Ratings to Decision Makers: Design for Good Practice. In M. Theall, & J. Franklin (Eds.) *Student ratings of instruction: Issues for improving practice. New directions for teaching and learning.* Number 43. San Francisco: Jossey-Bass.

Frey, P.W., (1973). Student ratings of teaching: Student ratings of teaching: Validity of several rating factors. *Science*, **182**, 83-85.

Frey, P.W. (1979). The Dr. Fox effect and its implications. *Instructional Evaluation*, **3**, 1-5.

Frey, P.W. (1976). Validity of student instructional ratings: Does timing matter? *Journal of Higher Education*, **47**, 327-336.

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

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

Gibbons, J.D. (1971). *Nonparametric statistical inference*. New York: McGraw-Hill.

Gillmore, G. M. (1973). *Estimates of reliability coefficients for items and subscales of the Illinois Course Evaluation Questionnaire*. Urbana, IL.: Illinois U., Office of Instructional Resources. 42p. ED082656.

Glass G. V.(1976). Primary secondary and meta-analysis of research. Educational Researcher, **5**

Glass, G.V. (1978).Integrating findings: The meta-analysis of research. In L.S. Shulman (Ed.). *Review of research in education* (Vol. 5). Itaska, IL: F.E. Peacock

Glass, G.V., McGaw, B., and Smith, M.L. (1981). *Meta-analysis in social research.* Beverly Hills, Calif.: Sage.

Gleser, L.J., & Olkin, I. (1994). *Stochastically dependent effect sizes.* In H. Cooper, & L.V. Hedges (Eds)*The handbook of research Synthesis.* New York: Russell Sage Foundation.

Gorsuch, R.I. (1983). *Factor analysis.* Hillsdale, N.J.: Lawrence Erlbaum.

Greenwald, A.G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, **82**, 1-12.

Grush, J.E., and Costin, F., (1975). The student as consumer of the teaching process. *American Educationa Research Journal*, **12**, 55-66.

Haladyna, T., & Hess, R. (1993, April). The detection and correction of bias in student ratings of instruction. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Hall, J.A., Rosenthall, R., Tickel-degnen, L., & Mosteller, F. (1994). Hypotheses and problems in research synthesis. In H. Cooper, & L.V. Hedges (Eds)*The handbook of research synthesis.* New York: Russell Sage Foundation.

Harman, H. H. (1976). *Modern factor analysis* (3rd rev. ed.). Chicago: University of Chicago Press.

Hartley, E. L., & Hogan, T. P. (1972). Some additional factors in student evaluation of courses. *American Educational Research Journal*, **9**, 241-250.

Hativa, N., & Raviv, A. (1993). Using a single score for summative evaluation by students. Research in Higher Education , **34**, 625-46

Hattie J. A., & Hansford B. C.(1984). Meta-analysis: A reflection on problems. *Australian Journal of Psychology*, **36**, 239-254.

Hazelton, A.E., (1980). *A study of the validity of student ratings of college teaching assessed on a criterion of student achievement in a first course in calculus.* Doctoral dissertation. University of Florida. (University Microfilms No. ADG81-05581).

Hedges, L.V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, **6**, 107-128.

Hedges, L.V. (1982a). Fitting categorical models to effect sizes from a series of experiments. *Journal of Educational Statistics*, **7**, 119-137.

Hedges, L.V. (1982b). Fitting continuous models to effect size data. *Journal of Educational Statistics*, **7**, 245-270.

Hedges L. V.(1986). Issues in meta-analysis. *Review of Research in Education.* **13**, 353-398.

Hedges L. V.(1987). How hard is hard science, how soft is soft science? The empirical cumulativeness of research. *American Psycholgist*, **42**, 443-455.

Hedges, L.V. (1994). Fixed effects models. In H. Cooper, & L.V. Hedges (Eds)*The handbook of research synthesis.* New York: Russell Sage Foundation.

Hedges L., & Olkin I.(1980). Vote-counting methods in research synthesis. *Psychological Bulletin*, **88**, 359-369.

Hedges, L.V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, Fl.: Academic Press.

Hill, P.W. (1984). Testing hierarchy in educational taxonomies: A theoretical and empirical investigation. *Evaluatiion Education*, **8**,181-278.

Hoffman, R.G., (1978). Variables affecting university student ratings of instructor behavior. *American Educational Research Journal,* **15**, 287-299.

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

Howard, G.S., & Maxwell, S.E. (1980). The correlation between student satisfaction and grades: A case of mistaken causation? *Journal of Educational Psychology*, **72**, 810-820.

Howard, G.S., & Maxwell, S.E. (1982). Do grades contaminate student evaluations of instruction. *Research in Higher Education*, **10**, 305-315.

Hunter, J.E., & Schmidt, F.L. (1990). *Methods of meta-analysis: Correcring error and bias in research findings*. Newbury Park: CA: Sage.

Hunter, J.E., Schmidt, F.L., and Jackson, G.B. (1982)*Meta-analysis: Cumulating research findings across studies*. Beverly Hills, CA. Sage.

Isaacson, R.L., McKeachie, W.J., Milholland, J.E., Lin, Y.G., Hofeller, M., Baerwaldt, J.W., & Zinn, K.L. (1964). Dimensions of student evaluations of teaching. *Journal of Educational Psychology*, **55**, 344-351.

Jackson, G. (1980). Methods for integrative reviews. *Journal of Educational Psychology*,_**73**, 444-449.

Jenkins, J. (1987). Implicit theories and semantic similarities: Reply to Marsh and Groves. *Journal of Educational Psychology*, **79**, 490-493.

Johnson, B.T. (1989). *DSTAT software for the meta-analytic review of research literature*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Johnson, D. W., Johnson, R. T., & Maruyama, G. (1983). Interdependence and interpersonal attraction among heterogeneous and homogeneous individuals: A theoretical formulation and a meta analysis of the research. *Review of Educational Research*, **53**, 5-54.

Jones, L.V., & Fiske, D. (1953). Models for testing the significance of combined results. *Psychological Bulletin*, **50**, 375-382.

Judd, C.M., Drake, R.A., Downing, J.W., & Krosnick, J.A. (1991). *Journal of*

Personality and Social Psychology, **60**, 193-202.

Kaiser, H.F., Hunka, S., & Bianchini, J.C. (1971). Relating factors between studies based on different individuals. *Multivariate Behavioral Research*, **6**, 409-414.

Kalton, G. (1983). *Introduction to Survey Sampling*. Beverly Hills, CA

Kavanagh. M.J., MacKinney, A.C., & Wolins, L. (1981). Issues in managerial performance: Multitrait -multimethod, analysis of ratings. *Psychological Bulletin*, **75**, 34-49.

Kelly, A. (1986).A method to the madness? Quantitative research reviewing. *Research in Education*, **35**, 25-41.

Kirk , R.E. (1982). *Experimental Design: Procedures for the Behavioral Sciences*. 2nd Edition. Belmot, CA: Brooks/ Cole Publishing Co.

Kish, L. (1965). *Sampling*. New York:Wiley.

Kishor, N. (1995). The effect of implicit theories on raters' inference in performance judgement: Consequences for the validity of students ratings of instruction. *Research in Higher Education*, **36**, 177-195.

Koch, S. (1981). The nature and limits of psycological knowledge: Lessons of a century qua "science". *American Psychologist*, **36**, 257-269.

Kraemer, H.C. (1983). Theory of estimation and testing of effect sizes: Use in meta-analysis. *Journal of Educational Statistics*, **8**, 93-101.

Kraemer H. C., & Andrews G.(1982). A non-parametric technique for meta-analysis effect size calculations. *Psychological Bulletin*, **91**, 404-412.

Kulik, J. A.(1983). Synthesis of research on computer-based instruction. *Educational*

Leadership, **41**, 19-21.

Kulik J. A., & Kulik C-L C.(1988, April). *Meta-analysis: Historical origins and contemporary practice*. paper presented at Annual Meeting of the American Educational Research Association. New Orleans, LA

Kulik, J. A., & Kulik, C. C. (1989). Meta-analysis in education. In *International Journal of Educational Research*, **13**, 221-340.

Kulik, J.A., & McKeachie, W.J. (1975). The evaluation of teachers in higher education. *Review of Research in Education*, **3**, 210-240.

Landman, J. T., & Dawes, R.M. (1982). Psychotherapy outcome: Smith and Glass; conclusions stand up under scrutiny. *American Psychologist*, **37**, 504-516.

Lang, A. (1904).

Larson, J.R., Jr. (1979). The limited utility of factor analytic techniques for the study of implicit theories in student ratings of teacher behavior. *American Educational Research Journal*, **16**, 201-211.

Lee, R., Malone, M., & Greco, S. (1981). Multitrait-Multimethod-Multirater analysis of performance ratings for law enforcement personnel. Journal of Applied Psychology, 66, 625-632.

Levinthal, L. (1975) Teacher rating forms: Critique and reformulation of previous validation designs. Canadian Psychological Review, **16**, 269-276.

L'Hommedieu, R., Menges, R. J., & Brinko, K.(1987, April). *Putting the back in meta-analysis: Issues affecting the validity of quantitative reviews*. Paper presented at the annual meeting of the American Educational Research Association.

Washington, DC.

Light R. J.(1987). Accumulating evidence from independent studies: What we can win and what we can lose. *Statistics in Medecine*, **6**, 221-228.

Light R. J., & Pillemer D.(1982). Numbers and Narrative: Combining their strengths in resaerch reviews. *Harvard Educational Review*, **52**, 1-26.

Light R. J., & Smith P. V.(1971). Accumulating evidence: Procedures for resolving contradictions among different research studies. *Harvard Educational Review*, **41**, 429-471.

Linn, R.L., Centra, J.A., & Tucker, L. (1975). Between, within, and total group factor analysis of student ratings of instruction. *Multivariate Behavioral Research*, **10**, 277-288.

Mansfield R., & Bussey T.(1977). Meta-analysis of research: A rejoinder to Glass. *Educational Researcher*, **44**, 3.

Marascuilo, L.A. (1971). *Statistical methods for behavioral science research*. New York: McGraw-Hill.

Marsh, H.W. (1980).The influence of student, course, and instructor characteristics on evaluations of university teaching. *American Educational Research Journal*, **17**, 219-237.

Marsh, H.W. (1982a). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *Educational Research Journal*, **19**, 485-497.

Marsh, H.W. (1982b). Validity of students rating of college teaching: A multitrait-multimethod analysis. *Journal of Educational Psychology*, **74**, 264-279.

Marsh, H.W. (1983a). Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics. *Journal of Educational Psychology*, **75**, 150-166.

Marsh, H.W. (1983b). Multitrait-multimethod analysis: Distinguishing between items and traits *Educational and Psychological Measurement*, **43**, 351-358.

Marsh, H.W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology*, **76**, 707-754.

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

Marsh, H.W. (1988). Multitrait-multimethod analysis In Keeves, J.P. (Ed.), *Educational Research, Methodology, and Measurement: An international Handbook* (pp. 570-580). New York: Pergamon.

Marsh, H. W. (1991a). Multidimensional students' evaluations of teaching effectiveness: A test of alternative higher-order structures. *Journal of Educational Psychology*, **83**, 285-296.

Marsh, H.W. (1991b). A multidimensional perspective on students' evaluations of teaching effectiveness: Reply to Abrami and d'Apollonia (1991). *Journal of Educational Psychology*, **83**, 416-421.

Marsh, H.W. (1993). The use of students' evaluations and an individually structured intervention to enhance university teaching effectiveness. American Educational Research Journal , **30**, 217-51.

Marsh, H.W. (1994). Comments to "Review of the dimensionality of student ratings of instruction" at the 1993 annual American Educational Research Association meeting by d'Apollonia, Abrami, and Rosenfield. *Instructional Evaluation and Faculty Development*, **14**, 13-19.

Marsh, H.W. (1995). Still weighting for the right criteria to validate student evaluations of teaching in the IDEA system. *Journal of Educational Psychology*, **87**, 666-679.

Marsh, H.W., & Cooper, T.L. (1981). Prior subject interest, students' evaluations, and instructional effectiveness. *Multivariate Behavioural Research*, **16**, 82-104.

Marsh, H. W., & Dunkin, M.J. (1992). Students' evaluations of university teaching: A multidimensional perspective. In J. Smart (ed.), *Higher Education: Handbook of Theory and Research,* Vol. VIII. New York: Agathon Press.

Marsh, H.W., Fleiner, J., & Thomas, C.S. (1975). Validity and usefulness of student evaluations of instructiona quality. *Journal of Educational Psychology*, **67**, 833-839.

Marsh, H.W., & Groves, 1987). Students' evaluations of teaching effectiveness and implicit theories: A critique of Cadwell and Jenkins (1985). *Journal of Educational Psychology*, **79**, 483-489.

Marsh, H. W., & Hocevar, D. (1984). The factorial invariance of student evaluations of college teaching. *American Educational Research Journal*, **21**, 341-366.

Marsh, H. W., & Hocevar, D. (1991). Multidimensional perspective on students' evaluations of teaching effectiveness: The generality of factor structures across academic discipline, instructor level and course level. *Teaching and Teacher Education*, **7**, 9-18.

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

Matt, G.E.(1989). Decision rules for selecting effect sizes in meta-analysis: A review and reanalysis of psychotherapy outcome studies. *Psychological Bulletin*, **105**, 106-115.

Matt, G.E., & Cook, T.D. (1994). Threats to the validity of research syntheses. In H. Cooper, & L.V. Hedges (Eds)*The handbook of research synthesis.* New York: Russell Sage Foundation.

Maxwell, S.E., & Howard, G.S. (1987). On the underdetermination of theory by evidence. *Journal of Educational Psychology*, **79**, 331-332.

McCallum, L.W. (1984). A meta-analysis of course evaluation data and its use in the tenure decision. *Research in Higher Education*, **21**, 150-158.

McGaw, J. (1988). Meta-analysis. In Keeves, J.P. (Ed.), *Educational Research, Methodology, and Measurement: An international Handbook* (pp. Z). New York: Pergamon.

McKeachie, W.J. (1979). Student ratings of faculty.: A reprise. *Academe*, **65**, 384-397.

McKeachie, W.J., Lin, Y.-J.,& Mann, W. (1971). Student ratings of teacher effectiveness: Validity studies. *American Educational Research Journal*, **8**, 435-445.

McKeachie, W.J., Lin, Y.-J.,& Mendelson, C.N.. (1978). A small study assessing teacher effectiveness: Does learning last? *Contemporary Educational Psychology*, **3**, 352-357.

Meehl, P.E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consuling and Clinical Psychology*, **46**, 806-834.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Reseracher*, **18**, 5-11.

Michigan State University. (1971). *Student instructional rating system: Stabillity of factor structure*. (SIRS Research Report #2). Office of Evaluation Services.

Miller, R.U . (1972). *Evaluating faculty performance*. San Francisco: JosseyBass.

Mintzes, J.J. (1977).*Field test and validation of a teaching evaluation instrument: The student opinion of teaching*. Windsor: Ontario. (ED# 146 185).

Morgan, W.D., & Vasche, J.D. (1978). An educational production function approach to teaching effectiveness and evaluation. *Journal of Economic Education*, 123-126.

Morsh, J.E., Burges, G.G., & Smith, P.N. (1956).Student achievement as a measure of instructor effectiveness. *Journal of Educational Psychology*, **47**, 79-88.

Mosteller, F., & Bush, R. (1954). *Selected quantitative techniques*. In G. Lindsey (Ed.), *Handbook of social psychology: Vol. 1. Theory and method*. Cambridge, MA: Addison-Wesley.

Murdock, R.P., (1969). *The effect of student ratings of their instructor on the student's achievement and rating*. Salt Lake City, Ut. (ERIC Document Reproduction Service No. ED 034 715).

Murphy, K.R., & Anhelt, R.L. (1992). Is halo error a property of the rater, ratees, or the specific behaviors observed? *Journal of Applied Psychology*, **77**, 494-500.

Murphy, K.R., & Balzer, W.K. (1989). Rater errors and rating accuracy. *Journal of Applie Psychology*, **74**, 619-624.

Murphy, K.R., Jako, R.A., & Anhalt, R. L. (1993). Nature and consequence of halo error: A critical analysis. *Journal of Applied Psychology*, **78**, 218-225.

Murray, H.G. (1980). *Evaluating university teaching: A review of research.* Toronto, Canada: Ontario Confederation of University Faculty Associations.

Murray, H.G. (1983). Low inference classroom teaching behaviors and student ratings of college teaching effectiveness. *Journal of Educational Psychology*, **71**, 856-865.

Murray, H.V. (1984). Impact of formative and summative evaluation of teaching in North American Universities. *Assessment and Evaluation in Higher Education*, **9**, 117-132.

Murray, H. G. (1991). Effective teaching behviors in the college classroom. In *Higher education: Handbook of theory and research* (pp. 135-172). NY: Agathon Press.

Murray, H.G., Rushton, J.P., Paunonen, S.V. (1990). Teacher personality traits and student instructional ratings in six types of university courses. *Journal of Educational Psychology*, **82**, 250-261.

Naftulin, D.H., Ware, J.E., & Donnelly, F.A. (1973). The Doctor Fox lecture: A paradigm of educational effectiveness. *Journal of Educational Psychology*, **71**, 856-865.

Nathan, B.R., & Lord, R.G. (1983). Cognitive categorization and dimensional schemata:

A process approach to the study of halo in performance ratings. *Journal of Applied Psychology*, **68**, 102-114.

Nathan, B.R., & Tippins, N. (1990). The consequences of halo "error" in performance ratings: A field study of the moderating effect of halo on test validation results. *Journal of Applied Psychology*, **75**, 290-296.

Neumann, L., & Neumann, Y. (1985). Determinants of students' instructional evaluation: A comparison of four levels of academic areas. *Journal of Educational Psychology*, **78**, 152-158.

Orpen, C., (1980). Student evaluation of lecturers as an indicator of instructional quality: A validity study. *Journal of Educational Research*, **74**, 5-7.

Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics*, **8**, 157-159.

Orwin, R.G. (1994). Evaluating coding decisions. In H. Cooper, & L.V. Hedges (Eds)*The handbook of research synthesis.* New York: Russell Sage Foundation.

Palmer, J.G., Carliner, G., & Romer, T. (1978). Leniency, learning, and evaluations. *Journal of Educational Psychology*, **70**, 855- 863.

Pambookian, H.S. (1972). The effect of feedback from students to college instructors on their teaching behavior. Doctoral dissertation. University of Michigan.

Pearson, K. (1933). On a method of determining whether a sample of given size n supposed to have been drawn from a parent population having a known probability integral has been drawn at random. *Biometrika*, **25**, 379-410.

Pedhazur, E.G. (1982). *Multiple Regression in Behavioral Research*. New York, NY.

CBS College Publishing.

Pigott, T.D, (1994). Methods for handling missing data in research synthesis. In H. Cooper, & L.V. Hedges (Eds)*The handbook of research synthesis.* New York: Russell Sage Foundation.

Pillemer D., & Light R.(1980). Synthesizing outcomes: How to use research evidence from many studies. *Harvard Educational Review*, **50**, 176-195.

Presby, S., (1978). Overly broad categories obscure important differences between therapies. *American Psychologist*, **33**, 514-515.

Price, D. J. de Solla (1975). *Science since Babylon.* New Haven: Yale University Press.

Prosser, M., & Trigwell, K. (1990). Student evaluations of teaching and courses: Student study strategies as a criterion of validity. *Higher Education*, **20**, 135-142.

Rankin, E.F., Greenmum, R., and Tracy, R.J., (1965). Factors related to student evaluations of a college reading course. *Journal of Reading*, **9**, 10-15.

Rasmussen, J.L., & Loher, B.T. (1988).Appropriate critical percentages for the Schmidt and Hunter meta-analysis procedure: Comparative evaluation of Type I error rate and power. *Journal of Applied Psychology*, **73**, 683-687.

Raudenbush, S.W. (1994). Random effects models. In H. Cooper, & L.V. Hedges (Eds)*The handbook of research synthesis.* New York: Russell Sage Foundation.

Raudenbush, S.W., Becker, B.J., & Kalaian, H. (1988). Modeling multivariate effect sizes. *Psychological Bulletin*, **103**, 111-120.

Reeves, S.S.(1989, March). *The effects of nonindependence on Hedges' test for homogeneity in meta-analysis research*. Annual meeting of the American Educational

Research Association. San Francisco, CA

Remmers, H.H., Martin, F.D., & Elliot, D.N., (1949). Are student ratings of their instructors related to their grades ? *Purdue University Studies in Higher Education*, **44**, 17-26.

Reynolds, D.V., & Hansvick, C. (1978, September). *Graduate instructors who grade higher receive lower evaluations by students*. Paper presented at the annual meeting of the American Psychological Association, Toronto, Ontario.

Rodin, M., and Rodin,B., (1972). Student evaluations of teachers, *Science*, **177**, 1164-1166.

Rosenfield, S., d'Apollonia, S., & Abrami, P. (1993, April). *The Dimensionality of Student Ratings of Instruction: Aggregation of Factor Studies.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Rosenthal, R. (1976). *Experimenter Effects in Behavioral Resrach*. New York: Irvington.

Rosenthal R.(1978). Combining results of independent studies. *Psychological Bulletin*, **86**, 1165-1168.

Rosenthal R.(1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*. **86**, 638-641.

Rosenthal R.(1984). *Meta-analytic procedures for social research.* Beverly Hills. CA:Sage Publicatiions.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper, & L.V. Hedges (Eds)*The handbook of research synthesis.* New York: Russell Sage Foundation.

Rosenthal, R. (1995).Writing meta-analytic reviews. *Psychological Bulletin*, **118**, 183-192.

Rosenthal, R., & Rubin, D.B. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences*, **3**, 377-386.

Rosenthal R., & Rubin D. B.(1986). Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological Bulletin*, **99**, 400-406.

Rubinstein,. J., & Mitchell, H. (1970).Feeling free, student involvement, and appreciation. *Proceedings form the 78th Convention of the American Psychological Association*, **5**, 623- 624.

Saal, F.E., Downey, R.G., & Lahey, M.A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, **88**, 413-428.

Sacks, H.S., Berrier, J., Reitman, D., Ancona-Berk, V.A., & Chalmers, T.C. (1987). Meta-analysis of randomized controlled trials. *New England Journal of Medecine*, **316**, 450-455.

Shaddish, W.R., & Haddock, C.K. (1994). Combining estimates of effect size.In H. Cooper, & L.V. Hedges (Eds)*The handbook of research synthesis.* New York: Russell Sage Foundation.

Slavin, R. E. (1983, April). *Meta-Nonsense: Misuse of Meta-Analysis in Educational Research*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, QC.

Slavin R. E.(1984). Meta-analysis in education: How has it been used?*Educational Researcher*, **13**, 6-15, 24-27.

Slavin R. E.(1986). Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. **15**, 5-11.

Slavin R. E.(1987). Best-evidence synthesis: Why less is more. *Educational Researcher,* **16**, 15-16.

Smith, R. A., & Cranton, P. A. (1992). Students' perceptions of teaching skills and overall effectiveness across instrucitonal settings. *Research in Higher Education*, **33**, 747-764.

Smith, M.L., & Glass, G.V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, **32**, 752-760.

Smith, M.L., & Glass, G.V. (1980). Meta-analysis of research on class size and its relationship to attitudes and instruction. *American Educational Research Journal*, **17**, 419-433.

Smith, M.L., Glass, G.V., & Miller, T.I. (1980). *The benefits of psychotherapy.* Baltimore: John Hopkins University Press.

Snyder, C.R., & Clair, M. (1979). effect of expected and obtained grades on teacher evaluation and attribution of performance. *Journal of Educational Psychology*, **68**, 75-82.

Solomon, D., Rosenberg, L., & Bezdek, W.E. (1964). Teacher behavior and student learning. *Journal of Educational Psychology*, **55**, 23-30.

Soper, J.C, (1973). Soft research on a hard subject: Student evaluations reconsidered, *Journal of Economic Education*, 22-26.

Sorge, D.H., & Kline, C.E. (1973).Verbal behavior of college instructors and attendant effect upon student attitudes and achievement. *College Student Journal*, **7**, 24-29.

Spector P. E., & Levine E. L.(1987). Meta-analysis for integrating study outcomes: A Monte Carlo study of it . *Journal of Applied Psychology*, **72**, 3-9.

Spencer, R.E., & Dick, W. (1965). *Course evaluation questionnaire: Manual of interpretation.* (Research Report No. 200). Urbana: University of Illinoise, Office of Instructional Resources.

SPSS, Inc. (1994). *SPSS 6.1 for Windows update*. Chicago, IL: SPSS Inc.

Stock, W.A.(1994). Systematic coding for research synthesis. In H. Cooper, & L.V. Hedges (Eds)*The handbook of research synthesis.* New York: Russell Sage Foundation.

Strube, M. J.(1986, April). *A general model for estimating and correcting the effects of non-independence in meta-analysis.* Paper presented at the Annual Meeting of the American Educational Research Association. April, San Francisco, CA

Strube, M.J. (1988). Some comments om the use of magnitude-of-effects estimates. *Journal of Counseling Psychology*, **35**, 342-345.

Strube M. J., & Hartmann D. P.(1982). A critical appraisal of meta-analysis. *British Journal of Clinical Psychology*, **21**, 129-139.

Sudman, S. (1976). *Applied sampling.* New York: Academic Press.

Sullivan, A.M., & Skanes, G.R. (1974). Validity of student evaluation of teaching and the characteristics of successful instructors. *Journal of Educational Psychology*, **66**, 584-590.

Summers, C.R.(1989, March). The derivation and preliminary application of the validity adjusted effect size. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Thompson, B. (1989). Meta Analysis of Factor Structure Studies: A Case *Study Example with Bem's Androgyny Method. Journal of Experimental Education,* **57,** 187-97.

Thorndike, E.L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, **4**, 25-29.

Thurstone, L.L. (1937). *The reliability and validity of tests.* Ann Arbor: Edwards.

Tippett, L.H.C. (1931). *The methods of statistics.* (1st ed.). London: Williams, & Norgate.

Tracz S. M., & Elmore P. B.(1985, August). *The issue of nonindependence in correlational meta-analysis.* Paper presented at the annual meeting of the American Statistical Association. LasVegas, NA.

Trzebinski, J. (1985). Action-oriented representations of implicit personality theories. *Journal of Personality and Social Psychology*, **48**, 1387-1397.

Turner, R.L., & Thompson, R.P. (1974, April). *Relationships between college students' ratings of instructors and residual learning.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Ware, J.E., & Williams, R.G. (1975). The Dr. Fox effect: A study of lecturer expressiveness and ratings of instruction. *Journal of Medical Education*, **5,** 149-156.

Wherry, R.L. (1951). *The control of bias in ratings: Factor analysis of rating item content.* Columbus, The Ohio State University Research Foundation, 1951,Sept.

Army, AGO, Personnel Research Branch, PRB Report No. 919.

Whitely, S.E., & Doyle, K.O. (1976). Implicit theories in student ratings. *American Educational Research Journal*, **13**, 241-253.

Whitely, S.E., & Doyle, K.O. (1979). Validity and generalizability of student ratings from between-class and within-class data. *Journal of Educational Psychology*, **71**, 117-124.

Widlak, F.W., McDaniel, E.D., & Feldhusen, J.F. (1973, February). *Factor analysis of an instructor rating scale*. Paper read at the annual meeting of the American Educational Research Association, New Orleans. LA. (ED# 079 324).

Wilkonson, B. (1951). A statistical consideration in psychological research. Psychological Bulletin, **48**, 156-158.

Wilson, T. (1987, April). *Pedagogical justice and student evaluation-of-teaching forms: A critical perspective.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Wolf, F.M. (1986). *Meta-analysis: Quantitative methods for research synthesis.* Beverly Hills, CA: Sage.

Wortman, P.M. (1994). Judging research quality. In H. Cooper, & L.V. Hedges (Eds)*The handbook of research synthesis.* New York: Russell Sage Foundation.

Yates, F., & Cochran, W.G. (1938). The analysis of groups of experiments. *Journal of Agricultural Science*, **28**, 556-580.

Zeller, R.A. (1988). Validity. In Keeves, J.P. (Ed.), *Educational Research, Methodology, and Measurement: An international Handbook* (pp. 322-330). New

York: Pergamon.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, **99**, 432-442.

APPENDIX 1

CODE BOOK FOR INSTRUCTIONAL CATEGORIES

Abrami and d'Apollonia (1990) showed that student rating forms differ greatly both in their inclusion of specific instructional categories and in the degree to which these instructional categories contribute in a "pure" form to the rating scores. Subsequently, they developed a coding schema from Feldman (1976, 1983, 1984,1989). This schema consisted in the following forty instructional categories grouped into six hierarchical categories. The bolded letters within parentheses are the codes that were used for the forty specific category in the tables.

1. Predispositions of instructor: This category includes the characteristics of the instructor that are present before the course begins. These characteristics often lead to the behaviours that take place once classes begin.

Personality Characteristics: The students are evaluating the instructor's

- personal appearance, peculiarities, health and attire **(AA)**;
- general personality characteristics such as maturity, irritability, confidence, paranoia, cynicism, prejudice, and tactfulness **(AB)**; and
- general attitudes **(AC)**. The attempt is first made to fit items into the other, more specific dimensions, and only if they do not fit elsewhere are they classified here.

Knowledge: The students are evaluating the extent to which the instructor demonstrated his/her

- specific knowledge of the course subject matter and its applications **(BA)**;
- general knowledge and cultural attainment beyond the course **(BB)**; and
- knowledge of pedagogy (eg. knowledge of students, student learning, and/or of instructional methods) **(BC)**.

Enthusiasm for and interest in subject, students, or teaching: The students are evaluating the extent to which the instructor demonstrated his/her enthusiasm, interest, concern or liking for

- the subject **(CA)**;
- students as students or as persons **(CB)**; and
- teaching, student learning or working with students **(CC)**.

Research productivity and reputation: The students are evaluating the research productivity and reputation of the instructor **(D)**.

2. Course Preparation and Organization: This category includes those behaviours that an instructor undertakes in preparing for a course, i.e. text selection, syllabus preparation, lecture preparation, etc. In a multi-section course these are often undertaken by someone in charge of the course and not necessarily the instructor. They are usually kept constant across sections and therefore one might expect a lower validity for this dimension.

Value of Course Materials and Supplementary Materials: The students are evaluating qualities such as relevance, value and clarity (but not difficulty) of the

- required course materials including textbooks, assignments, etc. **(EA)**; and
- supplementary materials (film, audio-visuals, etc.) **(EB)**.

Unless explicitly labelled "supplementary" such materials are considered required.

Course Preparation and Organization: The students are evaluating the extent to which the instructor decided on or prepared in detail the content, lecture, materials, tests or methods of instruction; and organized the topics or sequence of <u>activities</u> (class, course) logically or according to the text book. This dimension <u>only</u> relates to preparation not presentation. Any items that are ambiguous in terms of whether they relate to preparation or presentation are classified in presentation since students judge on the basis of presentation **(F)**.

3. Instructor Presentation Skills: This category captures the teacher's role as communicator. It includes the rhetorical behaviours that promote effective classroom presentations.

Captures students' attention: The students are evaluating the extent to which the instructor

- captured and maintained their attention by such means for example as stimulating their interest in the course, arousing their intellectual curiosity, as indicated by the attendance, the increased interest, outside reading and discussion and curiosity in and liking/enjoyment for the subject matter **(GA)**; and
- motivated students both to more effort and higher aspirations **(GB)**.

Course Objectives: The students are evaluating the extent to which the instructor communicates and impliments clear course objectives, criteria and deadlines **(H)**.

Instruction: The students are evaluating the extent to which the instructor

- uses appropriate instructional methods (e.g., lectures, discussion) and materials (e.g., textbook, A.V. materials) in class **(IA)**,
- delivers clear, concise, understandable and accurate instruction (lectures, laboratories, etc.) with appropriate explanations and summaries **(IB)**
- emphasises the relevance of the provided information including recent research **(IC)**; and
- encourages students to ask questions and responds to students' questions appropriately **(ID)**.

Presentation appropriate to audience and discipline: The students are evaluating the extent to which the instructor is responsive to the students' interests and experiences by being aware of student progress, by selecting an appropriate class level and pacing for instruction and by being willing to change when necessary **(J)**.

Elocutionary Skills: The students are evaluating the extent to which the instructor

- demonstrated skill in vocal delivery **(KA)**; and
- delivered instruction in an expressive, dynamic, dramatic or exaggerated manner **(KB)**.

4. Instructional Climate: This category captures the behaviours encompassed by the teacher's role as facilitator.

Classroom Management: The students are evaluating the extent to which the instructor

- managed the classroom environment in an authoritarian or participatory style (e.g., admitting errors, being open to criticism, sharing responsibility with students/resource personnel, determined content and evaluation) and in his/her classroom demeanour formal, dignified) **(LA)**;
- classtime **(LB)**; and
- issues of classroom control, e.g., noise, order, seating, calling on students and permissiveness of disruptions; supervision of tests and disciplinary actions when disruptions occurred **(LC)**.

Classroom Social Norms: The students are evaluating the affective climate of the classroom facilitated by the instructor. It includes the extent to which the instructor modelled, encouraged and achieved

- student interaction during classes (both student-student and student-teacher) **(MA)**;
- tolerance to a diversity of opinions, ideas and viewpoints **(MB)**;
- respect and consideration for others by such behaviours as listening without interrupting to others, not belittling or criticizing colleagues and students personally, treating others as equals, knowing students by name, being on time for lectures and appointments **(MC)**;
- a friendly classroom climate by being friendly, personable and by demonstrating a sense of proportion and humour **(MD)**.

Classroom Cognitive Norms: The students are evaluating the academic climate of the classroom facilitated by the instructor. It includes

- the standards of performance; and the extent to which the instructor modelled, encouraged and achieved low-level cognitive outcomes such as recall, recognition, knowledge, competence, reading/writing skills, professional skills and conventions by such means as asking drill-type questions and stressing memorization **(NA)**; and
- high-level cognitive outcomes such as independence of thought, reasoning, comprehension, analysis, meta-cognition, problem solving, evaluation and creativity by such means as asking open-ended questions, assigning original projects, applying knowledge to community problems *etc.* **(NB)**.

Availability and Helpfulness: The students are evaluating the extent to which the instructor

- was approachable, listened to student problems, and helped students **(OA)**; and
- was available outside of the classroom for assistance and extra-curricular activities **(OB)**.

Feedback: The students are describing the instructor's use of praise of good work, or discussion of error, review and feedback (frequency, positive/negative) and evaluating its effect on students such as constructive, helpful, producing fear **(P)**.

Workload: The students are describing the performance standards, the workload of the course or assignments (amount, pace, difficulty) and evaluating the appropriateness of the workload **(Q)**.

5. Evaluation: This category includes those items associated with the nature and fairness of the instructor's summative evaluation .

Evaluation: The students are evaluating the extent to which the instructor's tests were appropriate in terms of content, frequency, time allocation, weight, difficulty, validity and learning opportunity. They are also evaluating the instructor's fairness and consistency in grading **(R)**.

6. Global evaluations: this category includes those items associated with global assessments of the course and instructor.

Overall Course: The students are evaluating the overall worth of the course **(S)**
Overall Instructor: The students are evaluating the overall effectiveness of the instructor **(T)**
Overall Learning: The students are evaluating the overall perceived learning that took place including the achievement of short and long term objectives, the value of the learning and the achievement in grades **(U)**

APPENDIX 2
ITEMS PRESENT IN FACTOR STUDIES

Items are subdivided into forty instructional categories. Items in bold type were elininated at first pruning stage; items in italics type were eliminated at second pruning stage. Categories labelled with an asterix, were subsequently dropped; therefore all items within the category were also dropped.

Personal Appearance, Health, and Attire (AA):

Personal appearance.
Teacher very careless about dress.
Very pleasing appearance.
Wore wrinkled clothes.
Poor posture.
Not flashily dressed.
Wore same outfit daily.
Always immaculately clean.
Had excellent physical health.
Nice looking.
Physically handicapped.

Personality Characteristics and Peculiarities (AB):

Sense of proportion and humor.
Personal peculiarities.
Rate the instructor on the basis of poise and classroom mannerisms.
The instructor exhibited professional dignity and bearing in the classroom.
Enhanced presentations with humor.
Crabby.
Good natured.
Consistent.
A typical old maid (or bachelor) personality.
Immature emotionally.
Very prejudiced.

Considerate.
No sense of humour.
Tactless.
Wonderful sense of humour.
Cynical attitude repels students.
Did not inspire confidence.
Magnetic personality.
Tried to show off.
Well-rounded personality.
Self-reliance and confidence.
Always carried a yard stick.
Always in a hurry.
Restrained humor.

General Attitudes (AC):

Liberal and progressive attitude.
Had unethical attitudes.
Did not approve of extra-curricular activities.
Very biased politically.

Knowledge of Domain (BA):

Did not need notes.
Aware of scientific methods.
Up to date in his field.
The instructor has a thorough knowledge of his subject matter.

Knowledge of Teaching and of Students (BB):

No ability to handle students.

General Knowledge and Cultural Attainment (BC):

Admired for great intelligence.
Large background of experience made subject more interesting.

Enthusiasm for Subject (CA):

Interest in subject.
The instructor was enthusiastic when presenting course material.
Interested in all aspects of subject.

Enthusiasm for Teaching (CB):

The instructor seemed to consider teaching as a chore or routine activity.
Enthusiastic about teaching.
Enjoyed teaching class.
The instructor seemed to be interested in teaching.

Enthusiasm for Students (CC):

Sympathetic attitude toward students.
Rate the instructor on the basis of the instructor's apparent interest in working with students.
The instructor seemed to be interested in students as persons.
Interested in individual students.
Was the instructor considerate of and interested in his students ?
Always suspicious of students.
Afraid of students.
Lacked interest in students.
Kept up with student affairs.
Attended school functions as chaperon.
Helped students plan social events.

Research Productivity and Reputation (D):

Cooperative with other teachers.
Looked to for advice.
Had several college degrees.

Choice of Required Materials (EA):

The textbook was very good.

Readings and text valuable.

Assignments added to course understanding.

Did not go to trouble of making up assignments.

Homework assignments were helpful in understanding the course.

Overall I would rate the text book.

Rate the extent to which the text was a useful part of the course.

The assignments provided a valuable learning experience.

Choice of Supplementary Materials (EB):

The outside assignments for this course are just about the right length/somewhat too long/somewhat too short/much too long/much too short.

Had varied illustrations about topic covered.

Overall I would rate the supplementary readings/ excellent/ good / satisfactory/ fair/ poor.

More outside reading is necessary.

Preparation and Organization (F):

Course material was poorly organized.

Generally the course was well organized.

Rate the extent to which the instructor's lectures were well prepared.

The instructor was consistently prepared for class.

Rate the extent to which the instructor's lectures and other material were well prepared.

Absolutely no previous preparation for class.

Became confused in class.

Best organized of any class I have had.

Course material was disorganized and hindered understanding.

Came early to organize work.

The instructor was well prepared for each class.

The course was well organized.

Planned in advance.

Stimulation of Interest in the Course (GA):

Rate the instructor on the basis that she presents the material or content of this course in an interesting manner.

Rate the extent to which the instructor stimulated your interest in the course.

Increased subject interest.

Teaching style held your interest.

Rate the extent to which the instructor stimulated your interest in the course.

Do you now enjoy reading more than you used to?

Gained interest in American government.

Do more reading on topic.

Everyone attended regularly.

Knew how to hold attention in presenting materials.

Made lectures stimulating.

No attempt to make course interesting.

Students counted the minutes until class was dismissed.

I developed increased interest in the field.

I read in the field with active attention and enjoyment.

I did related readings that had not been assigned.

The instructor made this course as interesting as the subject matter wouldl allow.

My interest in the subject area has been stimulated by this course.

You were generally attentive in class.

I had discussions of related topics outside of class.

The instructor did not increase my interest in the subject matter.

Motivating Students to Greater Effort (GB):

Stimulating intellectual curiosity.

Rate the instructor on the basis that the teaching methods inspire stimulate or excite me intellectually.

Rate the instructor on the basis that she motivates me to think rather than just memorize material.

I developed motivation to do my best work.

Plan to take more courses.

Inspired many students to do better work.

Motivated students to work.

Instilled spirit of research.

Inspired class to learn.

The instructor's expectations for student performance were very low/ low / average/ high/ very high.

He continually emphasized grades.

He stimulated the intellectual curiosity of his students.

I have been putting a good deal of effort into this course.

You were interested in learning the course material.

The instructor motivated me to put forth a good effort.

He maintained definite standards of performance.

He stressed high quality work.

You felt that this course challenged you intellectually.

Objectives (H):

The direction of the course was adequately outlined.

Detailed course schedule.

The instructor was clear on what was expected regarding course requirements assignments exams etc.

Students always knew what was coming up next day.

There was considerable agreement between the announced objectives of the course and what was actually taught.

Made clear when assignments were due.

Gave good outline of course.

The instructor's objectives for the course have been made clear.

In my opinion the instructor has accomplished (is accomplishing) his or her objectives for the course.

Objectives stated and pursued.

He followed an outline closely.

*** Appropriate Use of Materials (IA):**

Lectures were too repetitive of what was in the textbooks.

I would rate the general quality of the lectures/excellent/very good/satisfactory/fair/poor.

Clarity of Instruction (IB):

Presentation of subject matter.

Rate the instructor on the basis of the organized class presentation.

Rate the instructor on the basis that she makes clear or simple the difficult ideas or concepts in this course.

The instructor did not synthesize ideas.

Rate the extent to which the instructor was successful in explaining the course material.

Presentations clarified material.

Presented clearly and summarized.

Instructor's explanations clear.

Presentation well prepared and integrated.

He explained clearly and his explanations were to the point.

Instructions not complete.

Covered subject well.

Made subject clear.

Presentations of materials especially good.

Students in constant state of uncertainty.

Learned and understood subject matter.

Did not understand teachers' lectures.

Skipped over aspects of subject not interested in.

The instructor summarized or emphasized major points in lectures or discussions.

The instructor appeared to relate the course concepts in a systematic manner.

The instructor's class presentations made for easy note taking.

Lectures facilitated note taking.

The instructor presented material in coherent manner emphasizing major points and making relationships clear.

The instructor was successful in making difficult material understandable.

Drawn out explanations.

Relevance of Instruction (IC):

The instructor's use of examples or personal experiences helped to get points across in class.

Good use of examples.

Contrasted implications.

Gave background of ideas and concepts.

Gave different points of view.

Discussed current developments.

Related subject to everyday life.

Has not changed course materials for several years.

Did not stick to subject.

Failed to cite applications of the subject.

To what extent did the instructor use examples or illustration to help clarify the material.

Answering Questions (ID):

Rate the instructor on the basis that he answers student's questions in a clear and concise manner.

Rate the extent to which the instructor responded effectively to student questions.

Encouraged questions and answers.

The instructor encouraged and readily responded to student questions.

Became angry when questions were asked.

No questions allowed between explanations.

Never understood questions asked by the students.

Students never asked teacher for advice.

The student had an opportunity to ask questions.

Monitoring Learning (J):

The instructor was skilful in observing student reactions.

Skilled at bringing out special abilities of students.

Worked with students individually.

Aware of individual differences in pupils.

Sensed when students needed help.

Used technical terms above students heads.

The instructor seemed to know when students didn't understand the material.

Vocal Delivery (KA):

Rate the instructor on the basis that she speaks clearly and is easily heard.

The instructor is clear and audible.

Speech very fluent.

Lectured inaudibly.

Occasional bad grammar detracted from speech.

Free from speech defects.

Knew how to talk about many things well.

Dramatic Delivery (KB):

Dynamic and energetic.

Talked with back to class.

Hard to believe.

Always remained seated.

Stood while teaching.

Management Style (LA):

The demands of the students were not considered by the instructor.

He decided in detail what should be done and how it should be done.

He was permissive and flexible.

Knack in dealing with all types of problems.

Never deliberately forced own decisions on class.

Classes always orderly.

Conducted class smoothly.

Never considered what class wanted.

Maintained a well organized classroom.

Weak in leadership questions.

Allowed freedom of speech.

Always willing to direct activities.

Did not observe rules and regulations.

Extreme dominance over class.

Flaunted authority.

Called on students alphabetically.

Seated students alphabetically.

Informal in classroom.

Teacher open to suggestions made by class.

Classes usually very quiet.

Demanded strict attention.

Students answers had to coincide exactly with teacher's.

Activities were orderly scheduled.

*** Time Management (LB):**

Rate the instructor on the basis that he presents class material at a rate or pace best for student learning.

The instructor used class time well.

The instructor generally presented the material too rapidly.

Course pace was (too slow-too fast).

Class time wasted.

Did not fill time up with trivial material.

For me the pace at which the instructor covered the material during the term was very slow/somewhat slow/just about right/somewhat fast/very fast.

The pace of the course was too slow.

Ideas and concepts were developed too rapidly.

Half of class time taken up with tests.

Classes ran overtime.

Had everything going according to schedule.

Supervision and Disciplinary Actions (LC):

Never had to discipline the students.

Made no move to discipline students.

Punishment was effective.

Interaction and Discussion (MA):

Encouraged class discussions.

Encouraged expression of ideas.

Students would not cooperate in class.

Group discussions encouraged.

Nothing accomplished in classroom discussions.

Very skilful in directing discussion.

Had class projects in which all students participated.

I would rate the overall value of class discussions/excellent/very good/satisfactory/fair/poor.

The instructor generally stimulated class discussion.

There was not enough student participation for this type of course.

Discussions were welcome.

Encouraged to participate.

Encouraged to express ideas.

Students shared ideas and knowledge.

The students often volunteered their own opinions.

Tolerance of Diversity (MB):

The instructor was open to other viewpoints.

Rate the instructor on the basis that he considers opposing viewpoints or ideas.

The instructor appeared receptive to new ideas and others' viewpoints.

Intolerant.

Presented both sides of every question.

Blinded to all viewpoints but own.

In this class I felt free to ask questions or express my opinions.

Students argued with each other or with the instructor not necessarily with hostility.

Presented both sides of the argument.

The instructor encouraged students to express opinions.

I increased my tolerance for unconventional approaches to truth.

In his class I felt free to ask questions to express my opinions and disagree.

Respect for Others (MC):

The instructor's attendance and punctuality have been consistently good.

He listened attentively to what class members had to say.

Irritated easily.

Very impatient with less able students.

Carried friendliness outside of classroom.

Built up confidence in students.
Gained class confidence very quickly.
Made students feel at ease.
Sarcastic if disagreed with.
Students did things to make teacher mad.
Always very polite to students.
Humiliated students.
Publically ridiculed some students.
Ridiculed students.
Very sincere when talking to students.
Did not ridicule wrong answers.
Belittled other teachers in own field.
Often arrived late for class.
Seldom if ever absent.
Not two faced.
Made point to call students by name.
Offered praise impartially where due.
Used brutal frankness.
Never criticized in an embarrassing manner.
Never criticized in a destructive way.
Never spoke harshly.
Often raised voice.
Teased students.

Friendly Classroom Climate (MD):

He was friendly.
Friendly towards students.
Discouraged students
Made students feel very insecure.
Very much at ease with the class.
Students often returned to chat with teacher.
Afraid of aggravating teacher.
Students afraid to disagree.

*** Low-level Cognitive Outcomes (NA):**

Increased knowledge and competence.
When people discuss topics in this field I can recognize when they are using good or poor arguments.
When a question comes up in conversation I can recall relevant information.
I discovered a variety if new points of view.
I developed significant skills in the field.
I developed familiarity with the conventions of the field.
Gaining factual knowledge (terminology classifications methods trends).
Developing specific skills competencies and points of view needed by professionals in the field most closely related to this course.
Learned factual information from course.

High-level Cognitive Outcomes (NB):

The instructor encouraged students to think for themselves.
The instructor encouraged the development of new viewpoints and appreciations.
Understand advanced material.
Ability to analyze issues.
I can think more coherently.
Developing a sense of personal responsibility (self reliance self discipline).
Discovering the implications of the course material for understanding myself, interests, talents, values, etc.).
Developing specific skills competencies and points of view that I can use later in life.
Intellectual curiosity in subject stimulated.
Gained general understanding of topic.
Encouraged students to think out answers.
The instructor raised challenging questions or problems for discussion.
I can understand relatively advanced presentations on the subject.
I can confront new problems and use general ideas or techniques and skills from the course to solve them.
I can analyze new complex material identify the major elements and interrelate the components.
I can organize and reorganize the elements of a complex problem and come out with a pattern not clearly there before.
I can identify and appraise judgments and values that enter into making

decisions in this field.

I became aware of implications of the subject matter in my own life.

I increased my concern for community projects related to the course.

I appreciate things I didn't appreciate before.

I developed my ability to marshal or identify main points or central issues.

I developed my ability to supply or identify data or appropriate information necessary to support or refute conclusions or generalizations.

I developed my ability to arrive at some kind of synthesis so as to produce a reasoned judgment.

I developed increased sensitivity and evaluative judgment.

I developed awareness of varying modes of confronting problems.

I developed the ability to function creatively in this field.

Learning fundamental principles generalizations or theories.

Learning to apply course material to improve rational thinking problem solving and decision making.

Learning how professionals in this field go about the process of gaining new knowledge.

Developing creative capacities.

Gaining a broader understanding and appreciation of intellectual cultural activity (music, science, literature, etc.).

Developing a skill in expressing myself orally or in writing.

Concern for Students (OA):

The instructor seemed genuinely concerned with student's progress and was actively helpful.

The instructor seemed to be concerned with whether the students learned the material.

Listened and willing to help.

Concerned about student difficulties.

The instructor maintained a generally helpful attitude toward students and their problems.

Too busy for talks with students.

Always listened to students troubles.

Very hard to talk to.

Availability (OB):

Rate the instructor on the basis of the ease at which an office appointment can be made.

Welcomed seeking help and advice.

Accessable to individual students.

Welcomed conferences.

The instructor was readily available for consultation with students.

Able to get personal help.

The instructor has not been readily available for consultation by appointment.

Feedback (P):

Instructor did not review promptly and in such a way that students could understand their weaknesses.

The instructor made helpful comments on papers or exams.

Rate the instructor on the basis of the information or feedback provide concerning the nature and quality of my work (considering all the factors involved in teaching this course).

Examination feedback valuable.

Reviewed test questions that majority of students missed.

Verbal or written comments on assignments have been constructive.

He criticized poor work.

Homework never graded.

Never returned tests to look over.

Tests or papers graded promptly.

Ignored wrong answers.

Rarely collected assignments.

Held reviews.

Throughout this course I have not been able to assess my progress and achievement.

He kept students well informed of their progress.

He told students when they had done a particularly good job.

*** Workload (Q):**

The scope of the course has been too limited; not enough material has been covered.

The instructor attempted to cover too much material.

Had to work hard.

Required a lot of time.

Heavy work load.

Course difficulty (easy to hard).

Course workload (light to heavy).

Workload/pace was difficult.

The material in this course has been beyond my previous academic experience.

He assigned a great amount of reading.

The instructor assigned very difficult reading.

He asked for more than students could get done.

Assignments not too long.

For my preparation and ability the level of difficulty for this course was very elementary/somewhat elementary/ about right/somewhat difficult/very difficult.

The workload for this course in relation to other courses of equal credit was/much lighter/lighter/about the same/heavier/much heavier.

The homework assignments were too time consuming relative to their contribution to your understanding of the course material.

You generally found the coverage of topics in the assigned readings too difficult.

The amount of work required for this course has been: very light/ light /average/heavy/very heavy.

Gave plenty of time for assignments to be completed.

Spaced homework evenly.

Difficulty of the subject matter of this course is very high/ high/ medium/low /very low.

Evaluation (R):

The types of test questions used were good.

Fair and impartial grading.

Grading reflected performance.

Grading indicated accomplishments.

Evaluation methods fair and appropriate.

Exams emphasized course content.

Tests indicated careful preparation.

Would not explain grading system.

Reputation for being stiff grader.

Examinations reflected important aspects of the course.

Overall I would rate the quality of the exams/ excellent/ very good /satisfactory/ fair/ poor.

Never gave written tests.

No systematic grading system.

Inadequate coverage of course work on examinations.

Graded on quantity not quality.

Frequent errors in grading tests.

Tests usually too long.

Failed a set percentage of students.

Gave surprise quizzes.

Fairness in grading.

The instructor told students how they would be evaluated in the course.

Rate the instructor on the basis of the fairness of the questions on this instructor's exams (or whatever main method is used to evaluate students).

Rate the instructor on the basis of the fairness of this instructor's grading system.

The examinations were too difficult.

Rate the extent to which the examinations tested your knowledge of the course material.

The evaluation system for this course was fairly applied.

Very poorly organized test questions.

Overall Course (S):

You generally enjoyed going to class.

Overall course rating.

How would you rate the over-all value of this course?

Have you enjoyed taking this course?

Students discouraged with course.

Has this course helped improve your reading skills?

Do you feel this course has helped you improve your grades?

Rate the overall effectiveness of the course.

Overall Instructor (T):

Rate the overall teacher's effectiveness.
General teaching ability.
Attitudes about teaching.
Would you recommend this course from this instructor?
Overall instructor rating.
Would you recommend this course from this instructor ?
How would you rate your instructor with respect to general (all-around) teaching a bility?
Overall evaluation of instructor.
Would like instructor as personal friend.
Learned a lot from teacher.
Students avoided this teacher's class.
Not qualified as a teacher.
Rate the overall effectiveness of the instructor.

*** Overall Learning(U):**

How much have you learned.
The contribution to my professional or career goals are excellent/ good/fair or average/somewhat poor/very poor.
The contribution to my general educational background is excellent/good/fair or average/somewhat poor/poor.
You have become more competent in this area due to this course.
How much have you been motivated by instructor.
My overall learning progress.
Learned something valuable.
How much did you learn from this instructor?7

APPENDIX 3

CODE BOOK FOR STUDY FEATURES

The operational definitions for the coded characteristics are given below: These are subdivided into seven sets: Book-keeping variables (used to manipulate the data), Methodological and Publication characteristics, Quality of Evaluation characteristics, characteristics of the Student Rating Form, characteristics of the Achievement Measure, Instructor and Student characteristics, and Course and Institutional characteristics. These are described briefly below:

Book-keeping Variables

Study Name: a six-character word consisting of the first four letters of the first author's name and the last two digits of the publication year.

Study Number: the rank of the studies (sorted alphabetically)

Between-multisection Course Identifier: the characteristic that distinguishes among validity coefficients based on both different students and different instructors reported in the same primary study; for example, physics courses, sections taught by experienced instructors.

Between-multisection Course Number: the rank of multisection courses sorted in the same order as presented in the primary study.

Within-section Identifier: the characteristic that distinguishes among validity coefficients based on different students but the same instructors reported in the same primary study; for example, male students, students using rationally deprived TRF, etc.

Within-section Number: the rank of the subsection sorted in the same order as presented in the primary study.

Validity Coefficient Identifier: the characteristic that distinguishes validity coefficients based on both the same students and the same instructor reported in the same primary study; for example, with ability control, TRF correlated to first achievement test.

Validity Coefficient Number: the rank of the validity coefficient sorted in the same order as presented in the primary study.

Interdependency Level: Degree of interdependence

1 if validity coefficient represents a study
2 if validity coefficient is used to compare sections with both different students and different instructors
3 if validity coefficient is used to compare subsections with the same instructors but different students
4 if validity coefficient is a repeated measure.

Methodological and Publication characteristics

Year: the decade in which the study was published (treated as a continuous variable).

1 30's
2 40's
3 50's
4 60's
5 70's
6 80's

Number of Sections: the number of sections in the multisection course

Computational Issues: the type of validity coefficient reported in the primary study

1 calculated from Spearman's rho, Kendall's tau or partial correlations.
2 calculated from averages
3 taken directly from study

Source of Study: the extent of peer review of study

1 thesis, study never published
2 report or paper presentation, study never published
3 study published in refereed journal

Quality of Evaluation characteristics

Timing: the timing of the evaluation relative to the formative or summative performance feedback

1 no information given
2 during the last week of the course or after the final exam
3 before the end of the semester

Administrator: the person(s) administrating the rating form

1 no information given
2 instructor(s)
3 other

Scoring Bias: the method of scoring the achievement test

1 no information given
2 by instructor
3 by researcher or person other than instructor

Test Bias: instructor knowledge of content of achievement test

1 no information given
2 prior knowledge
3 no prior knowledge

Group Equivalence: control for section differences in student ability

1 no information given
2 reported that no ability differences present (including student lack of knowledge of who is teaching course)
3 statistical control (validity coefficient has ability differences partialled out)
4 random assignment

Characteristics of the Student Rating Form

TRF Source: The origin of the rating form used to compute the validity coefficient:

1 local unstandardized TRF
2 TRF standardized over one department
3 TRF standardized over one or more universities

Response Scale: the response scale used in the rating form

1 no information given

2 forced choice

3 2 to 5 point Liekert scale

4 6 to 25 point Liekert scale

Student Anonymity: the degree of confidentiality of the evaluation

1 no information given or stated not anonymous

2 unsigned

Length: the number of items in the rating form

TRF Reliability: the reliability of the rating form categorized by a median split

1 no information given

2 less than .715

3 more than .715

Number of items: the number of items upon which the validity coefficient is based (factor length)

Completion Rate: the completion rate for the rating form

1 no information given

2 less than 80%

3 more than 80%

Diversity Index: the dimensionality of the set of items used to compute the validity coefficient determined using the Shannon-Wiener Diversity Index:

$$D = - \Sigma \; p_i{}^* \log_2{}^* \; p_i$$

where p is the proportion of items in each instructional category in the reported validity coefficient.

Factor Structure: the structure of the rating form used to compute the validity coefficients defined as the proportion of items contributing to each of the first order factors (PFAC) (see page 148).

Characteristics of the Achievement Measure

Source of Achievement Test: the origin of the test used to compute the validity coefficient

1 no information given or Local test
2 Test Item Bank
3 Standardized Achievement Test

Question Type: the type of criterion measure

1 no information given
2 skilled performance, oral test, project, or problems
3 essay
4 multiple choice or objective test

Frequency: the frequency of testing

1 no information given
2 multiple testing during the semester
3 a pretest and a posttest (the criterion measure)
4 only the one final, criterion measure

Length: the number of items on the final test categorized by a median split

1 no information given
2 short (less than 20 items)
3 long (more than 20 items)

Score Calibration: method of computing criterion measure

1 no information given
2 raw scores
3 weighted scores
4 standardized scores

Learning Criteria: the type of learning outcome that was used as a criterion

1 no information given
2 Affective or Attitudinal
3 General/ Factual
4 Ccmprehension, Skilled Performance or Mastery

Measurement scale: the grading scale of the final test

1 no information given
2 letter grade
3 numerical grade

Achievement Test value: the value of the final test categorized by a tertiary split

1 no information given
2 less than 40
3 between 40 and 100
4 more than 100

Achievement Reliability: the reliability of the rating form categorized by a median split

1 no information given
2 less than .70
3 more than .70

Instructor and Student characteristics

Instructor Rank: the rank of the instructor

1 no information given or mixed rank
2 graduate student or TA
3 faculty

Teaching Experience: the instructor's experience

1 no information given or mixed experience
2 new (less than 1 year)
3 experienced

Instructor Autonomy: the degree of autonomy that the instructor yielded over the course requirements and conditions

1 no information given
2 very little - instructor conducted discussions or tutorials but everything else standardized
3 common syllabus, text, assignments and instruction supervised or coordinated to maintain uniformity
4 no mention of attempt to maintain uniform instruction

Student gender:

1 no information given or mixed gender
2 female students
3 male students

Course and Institutional characteristics

Method of Instruction: primary method of instruction provided by instructors being evaluated

1 traditional lecture or presentation

2 lecture with discussion or tutorial, instructors providing discussions or tutorials

3 audio-visual or video presentation, with instructors providing discussion or tutorials

4 instructors providing laboratory or practical instruction

5 instructors providing recitation or drill and practice on problems

Teaching Duration: the hours per week of contact between instructor and students categorized by a median split

1 no information given

2 less than 3 hours

3 more than 3 hours

Course Discipline: the subject content of the course

1 Social Sciences, Psychology, and Language Arts including Communication

2 Physical and Natural Sciences, Mathematics and Computer Sciences

Course Duration: the duration of the course

1 no information given

2 one semester

3 two semesters (full year)

Season: the semester in which the course was given

1 no information given or more than one semester

2 Spring and Summer

3 January to May

4 September to January

Institution: the status of the institution (Carnegie Commission on Higher Education, 1976)

1 undergraduate (not granting doctorate)
2 graduate (granting doctorate)

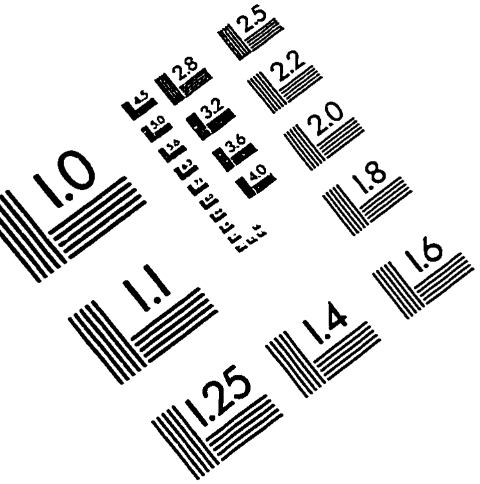
Multisection Class Size: the number of students in the course categorized by a tertiary split

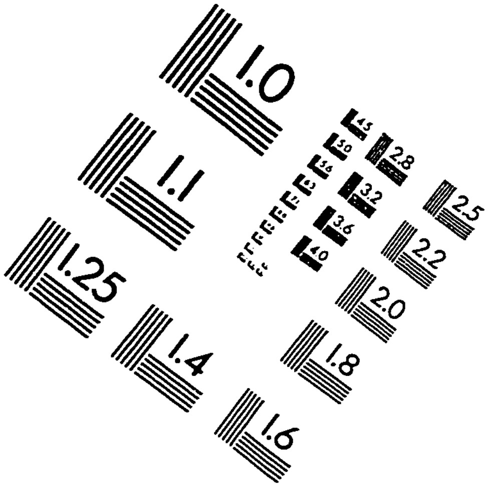
1 no information given
2 small (less than 400)
3 medium (between 400 and 600)
4 large (more than 600)

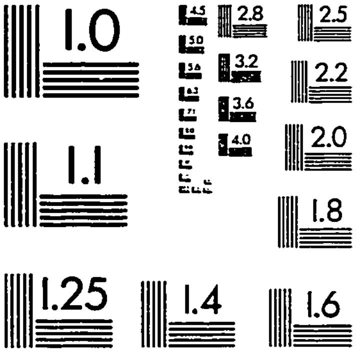
Section Size: the number of students in the section categorized by a tertiary split

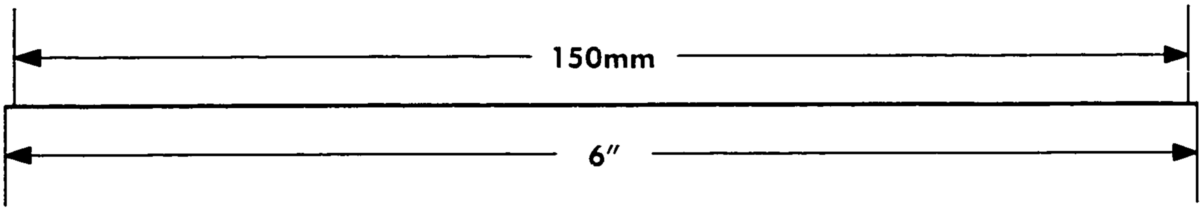
1 no information given
2 small (less than 17)
3 medium (between 17 and 23)
4 large (more than 23)

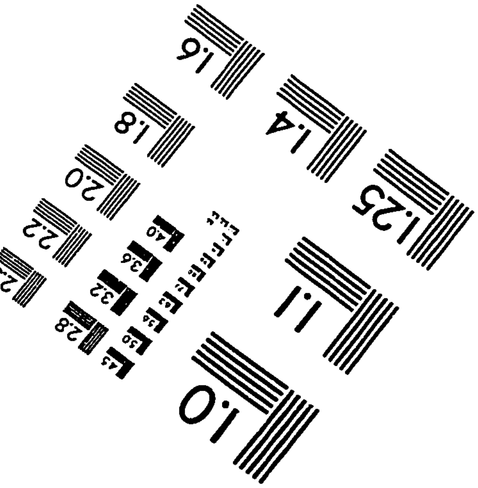
IMAGE EVALUATION
TEST TARGET (QA-3)

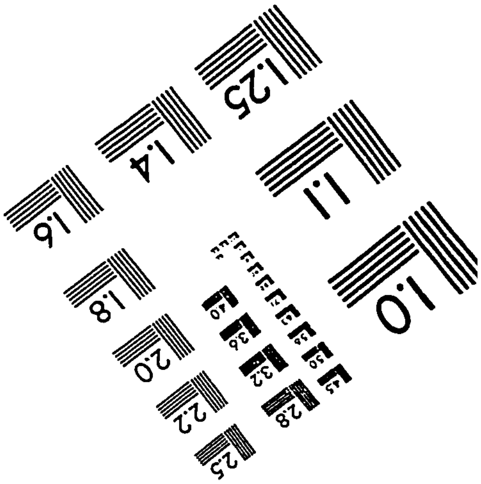












APPLIED IMAGE . Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989