A STUDENT-GENERATED QUESTIONNAIRE FOR THE EVALUATION OF TEACHING IN PSYCHOLOGY

by

Werner Johannes Müller-Clemm
B.Sc., University of Victoria, 1991

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF ARTS

in the Department of Psychology

We accept this thesis as conforming
to the required standard

[redacted]
Dr. C.W. Tolman, Supervisor (Department of Psychology)

[redacted]
Dr. L.K. Rosenblood, Departmental Member
(Department of Psychology)

[redacted]
Dr. J.N. MacGregor, Outside Member
(School of Public Administration)

[redacted]
Dr. J.C. McDavid, External Examiner
(School of Public Administration)

Supervisor: Dr. Charles W. Tolman

ABSTRACT

An important part of program evaluation research is to investigate methods for the production of reliable instruments with at least face validity. The evaluation of teaching in the university classroom is particularly challenging owing to its complexity and the variety of the activities involved, and also to the lack of generally accepted criteria for "good teaching." The most common way in which the problem has been approached has been to produce sets of descriptive statements or questions having to do with conformity to normally accepted standards of teaching practice and student reactions to teaching performance. Such rating instruments are commonly generated by the teaching or administrative staff.

The present research approached the problem within this format, but deviated from normal practice by seeking a way in which the questionnaire items could be generated by the students themselves. Of particular interest was to see whether the concerns expressed by the students regarding what items they felt ought to be included on a teaching evaluation questionnaire would differ significantly from those expressed by standard instruments. The research began by repeating an earlier attempt to proceed in this manner known as the "ECHO" technique (Schaefer, Bavelas, & Bavelas, 1980). Finally, a streamlined version of the earlier procedures was developed that appeared to deal more adequately with the problems of the reliability of instrument production.

A randomly selected group of students was asked to generate as many statements describing the value of courses and teaching as they wished. A second group was asked to categorize these items according to similarity of content, choose the most representative item of each category, and then rank the importance of each representative item.

The second procedure proved to be much to cumbersome. The focus was narrowed to only those items having to do with teaching performance. A further, less time-consuming, innovation was to provide a "discard" category into which items judged to be unimportant could be discarded. This procedure for sorting and ranking was then repeated by ten additional groups. The consistency of each item and its average ranked importance were calculated following the ten sorts. These provided the basic criteria according to which the final instruments were constructed.

Using a criterion of consistency of ten out of ten sorts, only three items in two categories remained. A reduction of the consistency criterion to eight out of ten sorts (the statistically reliable minimum based on binomial probabilities) yielded 15 items, representing nine distinct categories. These bore a remarkable resemblance to the item-content of the standard questionnaire currently used by the Psychology Department.

Applying the criterion of judged importance required a relaxation of the consistency criterion to five out of ten sorts yielding a set of 16 items where again, 9 distinct items with high rankings for importance were identified. The content of

these questions was somewhat different from those produced by the consistency (reliability) criterion.

Two nine-item questionnaires produced by this procedure, one (EC) based on a criterion of consistency of items and one based on judged perceived importance of item content (EI), were then presented to an independent sample of students from the same population, along with the standard departmental questionnaire (MD), reduced to nine items for comparability, to be preferentially rated for their expression of students' concerns. Using a paired-comparison procedure, the EC and MD forms showed no significant difference. The EI form, however, was judged as significantly more representative of student interests than either the EC or MD forms.

Implications for the practicability of student-generated teaching-evaluation questionnaires are discussed.

Examiners:

[redacted]

Dr. C.W. Tolman, Supervisor

(Department of Psychology)

[redacted]

Dr. L.K. Rosenblood, Departmental Member

(Department of Psychology)

[redacted]

Dr. J.N. MacGregor, Outside Member

(School of Public Administration)

[redacted]

Dr. J.C. McDavid, External Examiner

(School of Public Administration)

ACKNOWLEDGEMENTS

I am indebted to my entire committee for their participation and support. In particular, I wish to thank my supervisor, Dr. C. W. Tolman for his enduring patience and caring advice. As my mentor, he has instilled in me the importance of simplicity, critical analysis, and historical awareness. Through his example I have learned to fully appreciate A. N. Whitehead who insists that good teaching is embodied by the instructor whose teaching creates an "...atmosphere of excitement, arising from imaginative consideration, transform[ing] knowledge. A fact is no longer a bare fact: it is invested with all its possibilities" (1932/1955, p. 139).

DEDICATION

To my parents, Ruth and Berndt, who have always supported me in my endeavors. To Maria, for her loving tolerance and advice.

TABLE OF CONTENTS

Page

Page

Page

LIST OF TABLES

LIST OF APPENDICES

CHAPTER ONE

Introduction

The use of questionnaires for the student-rating of teaching performance, normally constructed by the faculty and/or the administration of a particular department, school, or university, is widespread in North America. The research and literature that deals with the construction of these questionnaires is equally abundant. For over half a century educators have wrestled with the issues surrounding the evaluation of teaching performance (Bendig, 1953a; Bendig, 1954; Blum, 1936; Remmers & Elliott, 1949). The evaluation of teaching performance is fraught with problems and conflicts. Experts, administrators, faculty, students, and business people all have a vested interest in the outcomes of research on the topic. The literature, in excess of 1300 articles and books (Cashin, 1988), supports that there is much interest concerning the issue of student-rating of teaching performance.

Literature Review

The current attitudes with regard to the rating or "evaluation" of faculty teaching performance by students can be described as a continuum. At the one extreme, the opinion amongst many university faculty seems to be that student-ratings of faculty-teaching performance are of little or no real evaluative merit[1] (Perry, 1990; Wright & O'Neil, 1992; Aleamoni, 1987; Costin, Greenough, &

---

[1] For detailed discussions of merit and worth see Guba and Lincoln (1981) or Scriven (1991). Simply put, merit is an estimate of intrinsic value (i.e., as related to an individual's teaching skills) and worth is an estimate of extrinsic value (i.e., the value of benefits - good publicity, income generated, etc. - that an individual represents to an employer) (Scriven, 1991).

Menges, 1971). From this perspective the value of such rating exercises is simply that of administrative "pap" generated to appease students and to fulfill bureaucratic needs. However, there is no adequate research to support this position.

The other extreme, supported by extensive research, represents the belief that student-ratings of faculty-teaching performance represent reliable and valid evaluations of instructional quality for both formative and summative purposes (Aleamoni, 1978; Aleamoni & Hexner, 1980; Aubrecht, 1979; Aubrecht, 1981; Cashin, 1988; Marsh 1984; Marsh, Overall, & Kesler, 1979)[2].

Between these two extremes lie various perspectives which embody varying degrees of agreement or disagreement with the above. There is, however, one viewpoint which summarizes this middle-ground in a representative way. This is manifested by those who accept formative, but are wary of summative uses of student-ratings of teaching performance (Boice, 1990-91; McKeachie, 1979). The formative/summative argument is, however, beyond the scope of the present research and serves only to differentiate the above perspectives. A position within this middle ground which is of interest to the present research is one which suggests that the evaluation of teaching performance employing student-ratings should be addressed through alternate methods and measures (Abbot, Wulff,

---

[2] Even at this end of the continuum most experts are still careful to wave a cautionary flag with regards to the unqualified and uncorroborated use of student-ratings of faculty-teaching performance for summative purposes. It is suggested that these ratings be employed in conjunction with other means of faculty evaluation (i.e., peer review, instructor self-ratings, etc.).

Nyquist, Ropp, & Hess, 1990; Cranton & Smith, 1990; Perry, 1990; Gitlin & Smyth, 1989; Doyle, 1983; Schaefer, Bavelas & Bavelas, 1980). These alternatives include the use of various forms of class interviews for collecting student opinions, evaluation feedback in the form of discussions between the students and instructors, different units of analysis (i.e., class means as opposed to individual ratings) and a standard method for the production of population-specific rating instruments as opposed to the use of standardized questionnaires.

Issues

Most research on the topic of student-ratings of teaching performance emphasizes and supports the reliability and validity of properly constructed rating instruments (Cashin, 1988; Aleamoni, 1987; Marsh, 1984; Cohen, 1981; Marsh, Overall, & Kesler, 1979). However, more recent research is beginning to question the generality of these findings (Abrami, d'Apollonia, & Cohen, 1990; Cranton & Smith, 1990; Murray, Rushton, & Paunonen, 1990).

What has been discovered is that, although moderately positive validity coefficients for global student-ratings have been reported, the results across studies are inconsistent and studies "have given insufficient attention to the quality and type of criterion measures studied..." (Abrami, d'Apollonia, & Cohen, 1990, p. 230). Abrami et al. (1990) also suggested that primary research must recognize that learning is multifaceted, differing in content, duration, and type. This point was also acknowledged by Marsh (1984) who stated that future research should recognize that "effective teaching and students' evaluations designed to reflect it

are multifaceted" (p. 707). In other words, "general research findings cannot...serve as the basis for individual teacher evaluations" (Guba & Lincoln, 1989, p. 61). Additionally, secondary effects such as changes in students' attitudes due to instruction, as well as the institutional context are factors which need to be considered. Accordingly, Abrami et al. (1990) "urge users of multidimensional rating forms away from the common practice of universally controlling for 'biasing' characteristics" (p. 230), since this eliminates valuable information necessary to an equitable evaluation.

Issues of validity in the area of student-ratings of teaching performance are complex. It is generally accepted that the most appropriate criterion of effective teaching is student learning[3]. Research studying this criterion has resulted in correlations (.22 to .50) which, although variable, indicate support for criterion validity (Cashin, 1988). However, these correlations are problematic due to the complex interrelations of student-learning variables with, for example, student characteristics, rather than being solely related to effective teaching characteristics (Cashin, 1988). Student-ratings of instruction have also been correlated with instructor's self-ratings (.20 to .69), ratings of administrators (.47 to .62), peers (.48 to .69), and alumni (.40 to .75) (Cashin, 1988).

---

[3] Scriven (1988) argues against this premise. He feels that personnel decisions about individuals cannot be based upon "statistical correlates of merit" since that kind of "stereotyping" is the same flaw found in racist or sexist practices (p. 12). Such decisions must therefore be based upon individual performance data.

Although the above research seems to have demonstrated a considerable amount of convergence, discriminability is also necessary to attain construct validity (Kerlinger, 1973). Theoretical concerns have therefore been raised by Cranton and Smith (1990) regarding claims of construct validity in student-rating research:

> The notions of implicit theories of interrelationships among instructor behaviors and the effect of the semantic similarity of questionnaire items raise serious concerns about the construct validity of student ratings (p. 208).

Another problem inherent in the rather narrowly focused, educational-evaluation research activities is fundamental. It concerns the identification (or lack thereof) of educational goals. In an essay titled "The Way Out of Educational Confusion", Dewey wrote that the growing complexity of the educational curriculum was creating educational confusion (1931).

This complexity has also led to confusion in the area of evaluating teaching effectiveness. In 1964 B. J. Biddle wrote that "we do not know how to define, prepare for, or measure teacher competence....For example, should a teacher's tasks be defined in terms of the ultimate goals of education or in terms of a specific effect upon the pupil?" (p. 3). Implied here, of course, is that the tasks of an instructor are extremely complex and interrelated on a variety of levels so as to make the identification of normative teaching criteria very difficult.

To this day, it appears that there is still a great deal of educational confusion. In a recent report (1992) published by the University of Victoria Task Force on Teaching (UVTFT), one of whose mandates was to make

recommendations on "the evaluation of teaching success" (p. 1), the reported "working definition of good teaching" was "any activity that enhances learning" (p. 2). In the "recommendations" section, this report stated that "evaluation policies and procedures should be designed to serve two purposes: to encourage the improvement of teaching (formative evaluation[4]), and to provide a clear and consistent basis (summative evaluation[5]) for rewarding excellent teaching..." (UVTFT, 1992, p. 8). Although multiple evaluation strategies were suggested and student-rating instruments were to be "well-designed", no concrete definitions of teaching, learning, or "well-designed" were offered.

Interim Summary

The "continuum" of perspectives on the student-rating of teaching performance described above represents the range of current attitudes on the topic. The opinions range from rejection to support and exemplify the continued disparity on the topic. One viewpoint within this controversy does not dismiss student-ratings, rather it suggests that alternate methods should be used in the evaluation of teaching. Although much of the research on the rating of teaching

---

[4] Formative evaluation is usually carried out during the course of a program etc. It has as its objective the improvement of a program, product, or person, etc. (Scriven, 1991). This type of evaluation tends to involve a larger proportion of the stakeholders, often in a more informal and cooperative setting. In university settings, faculty often design their own "in class" evaluations in order to improve the course and their own teaching skills.

[5] Summative evaluation is conducted at the conclusion of a program etc. (Scriven, 1991). In education, summative evaluation is used to make personnel decisions regarding faculty i.e., promotion, tenure, etc.).

performance by students supports the reliability and validity of properly constructed rating instruments, the results of recent meta-analyses show that generally, early research was often poorly structured and results were inconsistent.

Another basic issue is that of educational and institutional teaching objectives. The standardization of teaching, and therefore the evaluation of teaching, neglects the importance of context in the interaction between teaching and learning. The generalizable "facts" which are the result of this endeavor are not necessarily appropriate to the evaluation of individuals. Therefore the development of an alternative standard method for the generation of situation-specific student-rating questionnaires is the purpose of the present research. Such a method could more adequately incorporate the individual contextual differences inherent in the definition of "good teaching practices".

The "Echo" Answer

Research reported in 1980 by Schaefer, Bavelas and Bavelas identified two problematic assumptions upon which the, widely accepted, "standard" student-rating instruments are based. The first is that "generally, all students are essentially alike, within and across universities" and the second is that "all students are like the item writer, who is usually a professor and therefore a member of the population to be evaluated rather than the population who will respond to the items" (Schaefer, Bavelas & Bavelas, 1980). The acceptance of these assumptions raises various issues. Are the instruments reliable descriptors of students' criteria for rating instruction? Can the students' criteria be equated,

semantically with the standard questionnaires (i.e., do the items/instruments have equivalent meanings)? Can the indicators used by validation studies to justify student-ratings be considered true criterion (i.e., when peer ratings are correlated with student-ratings rather than with learning gains) (Scriven, 1988)?

A secondary, yet important, issue is the common use of the term "student evaluations" (Arreola, 1987; Cashin, 1988; Dalhousie University, 1991). It has been suggested that the use of the term "student ratings" rather than "student evaluations" would help to distinguish between the people who provide the ratings (students), and the people who interpret the data and perform the actual evaluations (administrators). This point further highlights the issue of semantics and the important, and perhaps too often neglected, role that language plays in the construction of student-rating questionnaires of faculty-teaching performance. The term "student ratings" will consequently be used in the present research.

Rather than attempting to construct yet another 'standardized' questionnaire, Schaefer et. al. (1980) suggested an alternative standard method for creating rating questionnaires. The "Echo" technique, originally developed by A. Bavelas (1942) and employed later by Barthol and Bridge (1968), was adapted in the attempt to construct a student-generated faculty-rating questionnaire (Schaefer, et. al., 1980). Although the Schaefer et al. (1980) study was remarkably different from the original "Echo" work done in 1942, the name "Echo" was retained by Schaefer et al. (1980) since the basic premise underlying the technique was preserved. This premise was conceptual, not procedural, and involved the

elicitation of information (questionnaire items, for example) from a specific population of people. The present research has also preserved this essential concept (although procedural changes have taken place) and consequently will refer to its method as the "revised Echo technique". The premise of the Schaefer et al. (1980) study was that the Echo technique would create a discipline-specific questionnaire which, when judged by students of that particular population, would be considered by them to be a more appropriate rating instrument than 'standard' questionnaires. Their results seemed to confirm their premise (Schaefer, et. al., 1980).

Idiographic/Nomothetic Approaches

The Echo technique employs an idiographic approach which makes "the explicit assumption that members of two different populations, such as students and professors, may not share the same equivalence class" [ie. the class of questionnaire items which encompasses for example, important aspects of teaching performance] (Schaefer, et. al., 1980). In contrast to the usual imposition of nomothetic principles in research, the idiographic approach emphasizes that "the classification of situations must be an integral part of any assessment procedure...[and] that such classification will have to be in terms of the individual's own phenomenology, not the investigator's" (emphasis added, Bem and Allen, 1974). Although their technique created a questionnaire which undeniably voiced the students' concerns, Schaefer et. al. (1980) admit that the

mere process of receiving student feedback regarding teaching performance does not automatically support its use in the evaluation of teaching performance.

The evaluation of an individual's teaching performance, therefore, can be approached from two directions. The idiographic, which insists that effective or "good" teaching is entirely dependent upon the individual's context, and the nomothetic, which believes that it is possible to have general laws of effective teaching which transcend the contextual "error" variances. The former approach lends itself very well to qualitative investigation. That is not to say that this would exclude quantitative data, on the contrary an individual student's score in a teacher's class may be a helpful indicator of that teacher's effectiveness. However, a reliance on just a few of such abstract criteria, without a detailed understanding of the entire context, would be highly inappropriate. In other words, the idiographic approach subsumes the results of nomothetically based methods, thereby allowing potentially important information (individual differences) to be considered.

Questionnaire Production

Although Schaefer et al. (1980) claim to have established content validity through a comparison of their instrument with other instruments, a much more fundamental issue was not addressed by them at all - that of consistency (or reliability) of questionnaire production. The terms consistency or reliability can be defined in a variety of ways, depending on the context of their usage. In the case of my research, these terms are interchangeable and are here defined as:

> the complex property of a series of observations, of a measuring instrument, or of the entire measuring process, that makes possible the obtaining of similar results upon repetition [emphasis added]; the degree to which such results may be predicted; the degree to which measurement is free from random influence (English and English, 1958, p. 456).

The most common definition of reliability, often linked with questionnaires or tests, deals with whether or not these instruments are answered in the same (reliable) way. In other words, to what extent does a group of people give the same response to a questionnaire (i.e., test-retest, split-half). This type of "response" reliability is not pertinent to the present research which is addressing issues of questionnaire production. The repeatability of results (Sidman, 1960) will therefore be referred to hereafter as the consistency of the production of a questionnaire in order to alleviate any confusion of terms.

An important characteristic of a good method is the ease with which it produces a result. In the present research the extent to which the Echo technique proved to be a smooth, rapid, and uncomplicated method for the production of student-generated questionnaires, was an important concern.

Consequently, answering the question "Can the Echo technique consistently and efficiently produce a set of student-generated items for the rating of teaching performance?", became the primary objective of my research. My goal was to improve upon and build a defensible case for using the Echo technique as a viable method of creating student-generated faculty-rating forms.

CHAPTER TWO

Method and Results

The purpose of this research was to establish whether or not the Echo technique is a viable method for the construction of student-generated questionnaires rating faculty teaching performance. The model for this technique was the "Echo" study by Schaefer, Bavelas, & Bavelas (1980).

Upon attempting to replicate the three major components of the original Schaefer et al. (1980) study (survey, sorting, and rating), methodological and procedural difficulties were encountered which prompted further pilot work. The methods and results of the five phases making up the present research will be described below in the following order: survey, pilot sorts, final sorts, questionnaire construction, and rating.

Phase I: Survey

Subjects

Subjects were undergraduate students enrolled in third and fourth year psychology courses at the University of Victoria (the total population was approximately 800 students). Fifty students were randomly selected from the population participated in the survey phase. Of these, forty-four responded and six either could not be contacted or declined to participate.

Materials

A minimal amount of materials was required for the survey phase. Ten by fifteen centimetre index cards, paper clips, and elastic bands were used. The

computer processing of the data was accomplished using Wordperfect 5.1 and Npstat 3.7.

Design and Procedure

Student class lists for third and fourth year psychology courses were obtained by the experimenter from the psychology department business office. Each enrolled student appearing on these lists was assigned a sequential number (eg., 1 to 1585). Numbers were selected from a random number table (Moses and Oakford, 1978). Whenever a selected number matched a name already chosen (due to many students being registered in more than one psychology class), a new number was selected. In this manner fifty different names were chosen for the sample.

The fifty students were contacted by telephone and were given verbal instructions (see Appendix A). These instructions were derived from the ones outlined by Schaefer et al. (1980). The most important points stressed in the instructions were: (a) that the scope of the questions should be limited to the evaluation of courses and faculty teaching performance, (b) that the subjects were to use their own, everyday language rather than attempting to be "academic," (c) that the subjects were to include a response format for each question generated, (d) that the subjects were to write as many or as few questions as they wished, (e) that the subjects were to use their intuition in voicing their true concerns, and (f) that their participation was strictly anonymous.

Subjects handed in their questions to the experimenter by prearranged appointment. Unsolicited anecdotal comments were (unsystematically) recorded by the experimenter from memory. The questions were then transcribed onto ten by fifteen centimetre index cards, one question and its response format per card. Although spelling mistakes were corrected, no grammatical, structural or other alterations were made.

Results

The survey portion of this research resulted in a pool of 289 items complete with response formats, 136 of which pertained to the rating of instruction. The survey phase had elicited on average over 6 items per subject on the evaluation of faculty and courses.

Verbal feedback from participants was generally positive regarding the research and its objectives. Recorded comments were made to the effect that students did not feel their participation in regular departmental student-ratings of faculty and courses had any impact or significance. Subjects believed that the current departmental rating form was meaningless to students ("A waste of time") in that it did not appropriately address their concerns. Interestingly, although doubt was voiced as to whether or not the written comments on the existing forms were read or acted upon ("They are just a way for us to vent..."), the majority, 197 (68 %) of the 289 items generated had, as part of their response formats, a "comments" option. Anecdotal comments are limited in their representativeness of the sample for two reasons:

1) subjects who did not have anything positive to say about my research or it's subject matter, may not have commented out of politeness, and/or

2) by recording the subjects' comments from memory I may have selectively recalled only the positive comments.

Phase II: Pilot Sorts

Initial testing of the original Echo sorting procedures led to further streamlining of these procedures to make the process a more expeditious one. After the first attempt at sorting had been completed it was clear that the subject matter of the items required a narrowing of focus to make the sorting easier. A second sort resulted in procedural changes which further reduced the difficulty and time of the sorting tasks.

First Pilot Sort

Subjects. Six volunteers registered in third and fourth year psychology classes participated in this procedure.

Materials. The 289 items generated in the survey phase, which had been transcribed onto index cards, were used. General office supplies such as pencils, rubber bands, and paper clips were also used.

Design and procedure. Subjects were seated at large tables in a comfortable seminar room and refreshments were provided. The procedure for sorting as described by Schaefer, Bavelas & Bavelas (1980) was followed (c.f. Bavelas, Bavelas, & Schaefer, 1978). The essential points of the procedure were:

a) The experimenter gave a brief outline of the purpose of the study followed by a synopsis of the sorting tasks.

b) Each subject received an equal share of the index cards (items), The cards had been shuffled by the experimenter prior to the start of each sort.

c) Task one: the subjects were asked to categorize the items into homogeneous categories.

d) Task two: the subjects were asked to take turns reading out loud the contents of one of their categories; subjects with categories which contained items corresponding to those items being read were to amalgamate their items with those of the person reading. This was done with the remaining categories (a miscellaneous category was created for those items which were not easily categorized).

e) Task three: the subjects were asked to go through each category, read the items out loud, and, as a group, decide if the categories were homogeneous. Items that were judged by subjects as heterogeneous to any given category, were re-categorized.

f) Task four: each subject was asked individually to go through each category and indicate, by a making a small pencil mark on the back, which of the items was the most representative of that category. Additionally, the subjects were asked to decide as a group, (1) which of the representative questions definitely should go on a

questionnaire, (2) which should definitely not go on a questionnaire and (3) which fell in between.

Results and problems. Although the procedure resulted in 24 categories, it was very problematic. The instructions were too lengthy and complex to be clearly understood and followed. The subjects were often "bogged down" by discussion and disagreement during the group activities. Subjects seemed to have difficulties during the categorization tasks, due to the great diversity of the items. The entire procedure lasted approximately five and one half hours. Three of the six subjects dropped out after two and three hours. The remaining three subjects were markedly fatigued at the conclusion of the study even though they resisted attempts by the experimenter to end the session.

Second Pilot Sort

From the results of the first pilot sorting procedure it was determined that changes needed to be made. It was decided that the time of the procedure might be reduced significantly by reducing the size of the pool of items. This reduction was accomplished by focusing the scope of the items more narrowly, i.e., through the elimination of all questions which did not have to do with the instructor. From the 289 items generated in the survey and used in the first pilot sort, 136 items (see Appendix B) specifically pertaining to the rating of faculty teaching performance were chosen for pilot sort two. This selection was carried out by the experimenter. The following criteria were employed: Any items containing the words "professor", "prof.", "teacher", "lecturer", or "instructor" were retained for

use. This effectively eliminated all items that did not deal with the teacher. Items referring strictly to examinations, textbooks, labs, teaching assistants, or assignments were thus excluded unless they were tied to the instructor in some way.

Subjects. Four volunteers registered in third and fourth year psychology classes participated. None had previously participated in the Echo study.

Materials. The reduced pool of 136 items resulting from the first sort were used. General office supplies such as pencils, rubber bands, and paper clips were also used.

Design and procedure. The conditions and instructions for this sort were the same as those used in pilot sort number one.

Results and problems. This second sorting procedure resulted in 19 categories. Although subjects still became "bogged down" in the group discussions, the experimenter perceived that the instructions seemed easier to follow, either because the pool of items was more homogeneous or because of greater intra-group agreement. This sorting procedure took approximately two and one half hours. One subject dropped out of the study after the first hour.

Summary of Pilot Sorts

At the conclusion of the pilot sorting procedures subjects in both pilot sorts were encouraged to provide suggestions regarding possible improvements to the study. A variety of positive comments were recorded regarding the general purpose of the study. Several constructive suggestions helped in the reformulation

of the sorting task and its instructions. The changes made were believed by the experimenter not to have significantly altered the sorting tasks other than to simplify them and make them less time-consuming. Briefly, the changes included the following aspects (see Appendix C for complete revised sorting instructions):

a) All of the sorting tasks were made into group tasks since the individual tasks in the pilot sorts required more interaction with the experimenter. Consequently, more time was consumed and the possibility of increased experimenter bias existed. Both of these aspects were undesirable.

b) Sorting groups were to be made up of three subjects each, since groups of this size seemed to interact more efficiently.

c) The objectives and instructions of each sorting task were streamlined by reducing the number and complexity of both.

Phase III: Final Sorts

From the results and experiences in the pilot sorting procedures, it was decided that the revisions made to the sorting procedure and task instructions were sufficient to warrant final implementation of the method.

Subjects

Ten groups of students were employed in the final sorting procedures (sorts 3-12). Each group contained three students who had not previously participated in the preceding phases of this study. All subjects were recruited from the population of third and fourth year psychology students. Participation was voluntary.

Subjects for the final sorting procedures were solicited in person by the experimenter from third and fourth year psychology classes and psychology department subject pool lists. The experimenter controlled the number of subjects in each session (3)[6]. Subjects were asked to sign up on a sign-up sheet (see Appendix D) and/or were contacted by telephone. They were given a brief description of the study, its objectives, and tasks. Subjects were then asked to indicate a day and time when it was convenient for them to participate. The subjects were otherwise self-selected into the sorting groups.

Materials

Use of the reduced pool of 136 items resulting from the first sort was continued. General office supplies such as pencils, rubber bands, and paper clips were used.

Design and Procedure

As in the pilot sorts, subjects were seated at large tables in a comfortable seminar room and refreshments were provided. Each sorting procedure consisted of four consecutive stages the details of which are described below. The subjects were given all information and instructions verbally, beginning with a brief review

---

[6] Each sorting group had three participants except for sort twelve. The lack of a third person meant that certain procedural changes had to be made. Task one: items were discarded only on consensus. Task two: all items retained were categorized on consensus. Task three: if a "tie" existed between two items chosen as most representative of a category occurred, both items would be accepted. Task four: rank ordering was done on consensus.

of the study, its purpose, and the subjects' tasks in it. This was followed by specific instructions at the beginning of each sorting task (see Appendix C).

Each group of subjects was allowed to work at their own pace. They were encouraged to relax and take breaks at their discretion. During the sorting procedure, a task's instructions were only repeated when the subjects asked to be reminded of them or when the experimenter determined that guidance was clearly required (eg., the task instructions were not being followed). The revised procedural strategies developed in the pilot sorts, were successfully applied in the final sorts (numbers 3 - 12).

The first sorting task involved the sorting of 136 items (see Appendix B) typed on individual index cards. The index cards were well shuffled before each sort and divided into three piles, one for each subject. The subjects were instructed to take turns reading an item aloud. They were asked to decide as a group, whether the item read should be placed in a "keep", "discard", or "miscellaneous" pile. The criteria for making these decisions was contained in the instructions (Appendix C). In order <u>not</u> to bias the way in which similar items were "kept" it was <u>very</u> important that the experimenter was attuned to the proceedings. It was extremely important that items were judge on their <u>own</u> merit, and were <u>not</u> compared to other similar items. The experimenter had to intervene when verbal signals were given by the sorting group which indicated that they were making such comparisons. Since the only means the sorting group used to communicate in determining whether or not to "keep" or "discard" an item was

verbal, this type of experimenter intervention was satisfactory. A typical example follows:

> Sorting group member reads an item and then says to the other group members: "Don't we already have an item like this? Experimenter intervenes and says: "Remember that during this task it is very important for you to judge each item on its own merit. Therefore, if an item is acceptable to you, no matter if you have already kept a similar one, keep it also. In other words, do not make any decisions based on comparisons between items, during this task. Unless you have any questions, please proceed."

The miscellaneous pile accommodated any initial lack of group consensus, limited discussion, and allowed this task to be completed relatively quickly and efficiently. Since the sorting task was the most time-consuming, the subjects were encouraged to limit their comments or discussions in order to allow them to remain focused on the task and expedite the procedure.

On the suggestion of subjects from the pilot sorts, the experimenter read out the items for the second task of the sort. Each item read was categorized under a temporary category label created by the participants of each individual sort. The category labels were not recorded as part of the data-gathering process since the goal was to generate a list of items, not categories. As the items were read they were placed either into an existing category, or a new category was created. During this task, subjects were still permitted to discard items which they

judged to be inappropriate. Subjects were reassured that, although the contents of each category were ideally to be homogenous, heterogeneous items would be "caught" during later tasks. The lack of rigidity of each task's instructions allowed the subjects to remain relaxed. This helped to reduce the time-consuming aspect of extensive discussions/debates inherent in forced decision-making situations and allowed the variability due to individual differences some flexibility. The items in the miscellaneous category were always read last and were also either categorized or discarded.

In the third sorting task the items were compared in the manner of a paired comparison, two at a time. Of the two compared, the item chosen as most representative of a category was compared against the next item in that category, and so on. If, during this task, an item was found to be heterogeneous to the category it was re-categorized. Discarding items was no longer an option during this task since all of the remaining items had been judged as "keepers" in stages one and two. Once all of the items from a category had been matched up against each other, the resulting "most representative" item was placed on top of the others and the category was secured with an elastic band.

The fourth stage required the subjects to rank-order the categories as to their relative perceived importance for the rating of faculty teaching performance, using the most representative items as their criteria. This was accomplished in two steps: (a) the items were grouped into three general clusters - "very important", "moderately important", and "least important", and (b) the items within

each of these clusters were then rank-ordered, again as to their perceived importance. All final decisions made by subjects within sorting groups regarding the fate of items, was accomplished through discussion. When subjects failed in the attempt to gain consensus, the decision was made by majority rule. Finally, the experimenter thanked the subjects for their participation. The results were coded and recorded by the experimenter, and the items shuffled for the next sorting group.

Summary of Method

During each sorting procedure the experimenter made notes regarding the procedure itself, identified potential or actual problems, recorded subjects' comments, and observed group peculiarities. At the conclusion of each sorting procedure subjects were asked if they had any comments or questions. No debriefing was necessary since the study employed no deceptive techniques and subjects were completely informed about the study prior to participation.

Results

Sorting procedures 3 to 12 employed the revised tasks and instructions that had evolved out the pilot work. A summary of the sorting results in tabular form are depicted in Table 1 by sort number, number of participants, duration of the procedure (in minutes), number of categories created, number of items "kept", and number of items "discarded".

Each sorting group began with the same pool of 136 items (see Appendix B). Against the criterion of 100% survival, only three items in two categories

Table 1

Summarized Results of Final Sorting Procedures

Sort #	n	Time (min.)	# of Categories	Number of Items: "Kept"	Number of Items: "Discarded"
3	3	115	15	64	72
4	3	90	17	46	90
5	3	100	17	50	86
6	3	105	16	34	102
7	3	100	15	34	102
8	3	95	18	73	63
9	3	110	29	111	25
10	3	105	22	64	72
11	3	85	16	40	96
12	2	95	18	50	86

were agreed upon by all subjects. These were: (1) "Did the professor clearly explain examination formats?", (2) "Would you take another course with this instructor?", and (3) "In the future would you take a course taught by the same instructor?".

Phase IV: Questionnaire Construction

After analyzing the results of the sorting procedures the decision was made to employ three questionnaires for the rating phase. These are described below. The reason for using these three forms was to have a comparison of a "standard" form, with two forms generated by the Echo technique, one being comprised of items based on a criterion of statistical consistency (or reliability) across sorts, and the other based on the subjects' perceived relative importance of items.

The Echo-consistent (EC) (see Appendix E) questionnaire was constructed by lowering the consistency criterion to 80%[7]. The justification for this criterion is based upon a binomial expansion test $p. = 0.03$ (using the empirical probability calculated for these data (average number of items "kept" divided by total number of items, $p. = .42$). The Echo-important (EI) questionnaire, shown in Appendix F, was constructed by (a) using the average rank-ordering of perceived importance results (the sum of an individual item's ranking in each of the ten final sorts, if and when "kept", divided by the number of times the item was "kept"; the lower

---

[7] The number of items consistently "kept" (retained) from sort to sort, as opposed to those items which were "discarded". At the 80% consistency level, eight out of the ten sorting groups would have to have "kept" the items (also see "Results" section p. 36).

the resulting score, the higher the average rank-order, see Table 2 (p. 31) for example). This criterion was identified by the experimenter as a possible source for a list of qualitatively different student-generated items. The average rank-order of perceived importance criterion was then combined with (b) lowering the consistency criterion to 50%[8]. The 50% criterion was judged by the experimenter as a reasonable (although not statistically supported) cut-off for the number of times an item was "kept" by sorting groups (yielded a "short list" of items similar in length to the 80% criterion). In combination these two criteria virtually assured that only the rank-order-of-importance determined items were selected for the EI questionnaire. The more an item was "kept" by sorting groups, the poorer was its mean rank-order (the exception was unusual) since it is more likely to have had variable rankings. Conversely, the less an item was "kept" by sorting groups, for example in the case of the relaxed 50% criterion, the better its mean rank-order of perceived importance rating.

In the case of multiple items within a single category, the most representative item of a category was chosen by the experimenter using the number of times (across sorts) that an item was chosen by the subjects in the sorts as "most representative". This criterion was used in the construction of both the EC and EI questionnaires.

The questionnaire used in the Department of Psychology at the University of Victoria (UV) (see Appendix G) for student-ratings of faculty and courses was

---

[8] See footnote 7 and "Results" section p. 38.

used in the construction of the Modified Departmental questionnaire (MD) (see Appendix H). In order for the ratings of the departmental form to be maximally based upon the merit of the items rather than be potentially prejudiced by recognition, the UV items were (a) judged by the experimenter as to their similarity of meaning with those of the Echo-consistent form (see Appendix E), in order to arrive at an equivalent number and type of items and, (b) the declarative statements on the departmental questionnaire were modified into question form so that they had the same format as the Echo items. In addition, only those UV items which pertained specifically to the rating of faculty teaching performance were retained for use in the MD form.

The three questionnaires were then compared by the investigator as to their relative semantic equivalence (i.e., items judged by the experimenter to have similar meanings). The reduction of items for the EC (15 to 9) and EI (16 to 9) forms, were decisions made by the experimenter based upon the number of times an item was chosen by sorting groups as most representative of its category (e.g., if one item was chosen twice, and another item was chosen three times as most representative, the latter item would be selected for use on a given "Echo" questionnaire).

Results

Upon lowering the criterion to 80% consistency a list of 15 items in 9 distinct categories was produced. These questions became the basis for the first Echo-generated questionnaire EC (see Appendix E).

A second list of "Echo" items (EI) (see Appendix F) was generated using items whose average 'rank-order-of-perceived-importance' was the highest. This criterion was combined with the consistency criterion which was additionally reduced to 50%. This produced a list of 16 items in 9 distinct categories.

The average rank-order of perceived importance for an item was calculated by summing the item's rankings across sorts 3 to 12 inclusive. This total was then divided by the total number of times that the item was "kept" as opposed to "discarded". The resulting figure was the item's mean rank-order value. Examples are listed in Tables 2, 3, 4, 5, and 6. A t-test was preformed on the items' resulting mean rank order values which confirmed that there was a significant difference between the means of EC and EI ($t = 2.75$, $df = 8$, $p. = 0.025$). This result supports the earlier assertion that items selected by subjects to be important were virtually assured of being chosen for the EI questionnaire since they usually had the better mean rank order values. Conversely, items which were selected for the EC questionnaire using the consistency criterion, were less likely to be chosen for the EI form since these items usually had poorer mean rank order values.

Upon comparing the three questionnaires as to their semantic equivalence, 8 of 9 items were found to convey similar meanings between MD and EC (for example: "Would you take another course with this instructor?", and "In the future would you take a course taught by the same instructor?").

Table 2

Example of Mean Rank-Order-of-Items Calculation:

Item not Chosen for a Questionnaire by any Criteria

Item # 1: "DID THE PROFESSOR MAKE THE MATERIAL TO BE TAUGHT AS INTERESTING AS POSSIBLE?"

				Sort #							
										Sum	Mean
3	4	5	6	7	8	9	10	11	12	R/O	R/O
	Rank-Order* (R/O) of Item in Sort										
14	1	17	X	7	9	9	X	X	3	60	8.57

* An "X" indicates that the item was not "kept" by that sorting group.

Table 3

Example of Mean Rank-Order-of-Items Calculation:

Item with 100% Consistency Rate

Item # 17: "DID THE PROFESSOR CLEARLY EXPLAIN EXAMINATION FORMATS?"

Sort #										Sum	Mean
3	4	5	6	7	8	9	10	11	12	R/O	R/O
Rank-Order* (R/O) of Item in Sort											
1	2	3	4	10	8	1	10	10	7	56	5.60

* An "X" indicates that the item was not "kept" by that sorting group.

Table 4

Example of Mean Rank-Order-of-Items Calculation:

Item Having Attained 80% Consistency Chosen for EC Questionnaire

Item # 33: "DOES THE INSTRUCTOR STAY ON TOPIC AS DEFINED BY THE COURSE OUTLINE?"

Sort #											
										Sum	Mean
3	4	5	6	7	8	9	10	11	12	R/O	R/O
Rank-Order* (R/O) of Item in Sort											
4	X	1	15	3	1	2	9	X	10	45	5.63

* An "X" indicates that the item was not "kept" by that sorting group.

Table 5

Example of Mean Rank-Order-of-Items Calculation:

Item with Best Mean Rank-Order but not Chosen for a Questionnaire by any Criterion

Item # 7: "DO YOU FEEL THAT THIS PROFESSOR HAS SUFFICIENTLY BROKEN DOWN THE MATERIAL OF THE COURSE IN A MANNER THAT RENDERS IT UNDERSTANDABLE?"

Sort #										Sum	Mean
3	4	5	6	7	8	9	10	11	12	R/O	R/O
Rank-Order* (R/O) of Item in Sort											
X	X	X	X	X	X	6	4	X	1	11	3.67

* An "X" indicates that the item was not "kept" by that sorting group.

Table 6

Example of Mean Rank-Order-of-Items Calculation:

Item Having Attained at Least 50% Consistency, Chosen for EI Questionnaire

Item # 96: "HOW PREPARED WAS THIS INSTRUCTOR?"

Sort #										Sum	Mean
3	4	5	6	7	8	9	10	11	12	R/O	R/O
Rank-Order* (R/O) of Item in Sort											
4	X	X	1	7	6	5	1	5	X	29	4.14

* An "X" indicates that the item was not "kept" by that sorting group.

When EC was compared to EI, 5 of 9 items were judged by the experimenter to have similar meanings. When MD was compared to EI, 6 of 9 items were judged by the experimenter to have similar meanings (see Table 7 for a summary of these results).

Phase V: Rating

As in the Schaefer et al. (1980) study, a validation procedure was performed. The main difference, however, was that the present research did not require subjects to rate faculty-evaluation questionnaires from outside of the psychology department or of the University of Victoria. The reasons were that: (a) the Echo technique led to the generation of two distinctly different questionnaires based upon two diverse approaches to the interpretation of the sorting results, and (b) until the consistency of questionnaire production was established within the discipline of psychology, it was illogical and methodologically inappropriate to move outside the discipline.

Subjects

The rating phase employed students registered in four sections of a third year psychology methods class (N = 145). Participation was voluntary and anonymous (only one potential subject declined to participate).

Materials

The rating phase involved comparisons of the three questionnaires constructed by the experimenter.

Table 7

Summary of Results: Semantic Equivalence

Item Number* in DP:	Semantic Equivalent Item Number* in:	
	ES	EP
1	1	1
2	2	2
3	X	7
4	5	X
5	6	6
6	3	X
7	4	X
8	8	8
9	9	9

* Item numbers are not marked on questionnaires (Appendices E,F,H) therefore the item number equals the order of presentation on the respective questionnaires.

Note: "X" denotes no item was judged to have the same meaning.

Design and Procedure

The questionnaires EC, EI, MD (see Appendices E, F, and H,) were compared in four sections of a psychology methods course at the University of Victoria (Psychology 300b, S01, S02, S03, S04). The forms were presented using a paired-comparison procedure with the questionnaires balanced for order of presentation (one form on top of the other) in the following six combinations: MD with EC, EC with MD, EC with EI, EI with EC, MD with EI, and EI with MD.(see Appendix I for verbal instructions and Appendix J for rater instruction/response cover sheet).

Students were told that their participation was voluntary. Participants were asked to take five minutes to read through the instructions and the attached questionnaires for the rating of faculty teaching performance. The forms were distributed to the entire class by the experimenter to ensure an even distribution of the six combinations of forms. To assure anonymous participation, the subjects were asked to cooperate in turning their forms face-down for collection whether or not they had participated.

The cooperation of faculty for the rating procedure was secured prior to its execution. The procedure was carried out mid-way through the school term. Class time was always used, either at the start or at the end of the class period.

In the present research, the rating procedure was intended to determine whether or not the revised Echo technique had generated items that responded more appropriately to students' concerns. The use of a multi-section class is

beneficial since it (a) eliminates the chance of a person participating more than once and, (b) in this case provided a relatively good representative sample of the population due to the fact that the course is required for psychology majors.

Results

The resulting data from the comparisons of the three questionnaires are summarized in Table 8. Chi-square analyses, corrected for continuity, were employed. Results indicated that at the first level of comparison: MD and EC were not judged significantly different (X^2 = 0.00, df = 1, N = 49, p. = 1.00) and EC and EI were not judged significantly different (X^2 = 1.69, df = 1, N = 48, p. = 0.19), whereas EI was judged as preferred against MD (X^2 = 4.69, df = 1, N = 48, p. = 0.03).

At the second level of comparison, chi-square analyses showed that: MD compared with both EC and EI were not judged significantly different (X^2 = 2.64, df = 1, N = 97, p. = 0.10) and EC compared with both MD and EI were not judged significantly different (X^2 = 0.66, df = 1, N = 97, p. = 0.42), whereas EI, when compared with both MD and EC was judged as the preferred questionnaire (X^2 = 6.51, df = 1, N = 96, p. = 0.01). There was no significant order of presentation effect (X^2 = .44, df = 1, N = 145, p. = .51) (see Table 9).

Table 8

Summary of Rating Component Results

Questionnaire Combinations	Class Section S01	S02	S03	S04	Cumulative Total
MD/EC	7/9	10/6	1/8	6/2	24/25
MD/EI	4/10	7/8	3/9	2/5	16/32*
EC/EI	8/8	4/11	3/6	4/4	19/29
MD/EC and EI	11/19	17/14	4/17	8/7	40/57
EC/MD and EI	17/15	10/21	11/7	6/10	44/53
EI/MD and EC	18/12	19/11	15/6	9/6	61/35**

* $p < .05$

** $p < .01$

Table 9

Summary Data for Calculation of Order Effect

Order of Presentation	Class Section S01	S02	S03	S04	Cumulative Total
First	22	24	11	11	68
Second	24	22	19	12	77

CHAPTER THREE

Discussion

Much of the current literature on the subject of student-rating of faculty teaching performance supports the view that students can give valid and reliable ratings. Questions and criticisms of previous research have been raised regarding the "common practice of universally controlling for 'biasing' characteristics" in the use of multi-dimensional rating forms (Abrami et al., 1990, p. 230). It was found that "characteristics that predict the relationship between student perceptions of teaching and instructor effects on learning vary with the aspect of teaching being investigated" (Abrami et al., 1990, p. 230). Therefore, this recent research suggests that when evaluating teaching performance an attempt must be made to take into account the individual context and interrelations of the instructor with his or her teaching environment, as well as the various perspectives of stakeholders.

The emphasis of the present research was as follows:
(a) to investigate the possibility that, when asked to construct a quasi-idiographic rating instrument, students' concerns about the quality of teaching would be different from those expressed by the faculty and administration through the "standard" questionnaires and, (b) that the revised Echo technique is an appropriate method for the production of such an instrument.

The revised Echo technique established itself as a relatively efficient method for the gathering of student feedback. Students in the survey phase were

enthusiastic and responded well to the instructions given to them. In the sorting component, tasks and their instructions were generally well received and subjects seemed to enjoy participating. The rating component was also successful in eliciting good participation. Recorded comments from subjects in all phases of this research indicated that the participants felt that the research was giving them a genuine opportunity to give their input regarding the evaluation of faculty teaching performance.

Two procedural changes are suggested for any future replication of this research: 1) The "categorization" task within the sorting procedure can be eliminated. This portion of the sorting procedure proved to be superfluous since the categories created were never the subject of investigation. Consequently, an intensive "keep"/"discard" procedure combined with the rank-ordering of "kept" items should not change the outcome of the sorting component, other than to speed it up; 2) Experimenter "judgements" should be eliminated whenever possible and should be replaced by judgements made employing population samples.

One of the main underlying questions of this research is whether or not students can consistently choose items, generated by a sample of the same population, for the rating of faculty teaching performance. The most rigorous answer (at 100% survivability) is "that they can but in a very limited way" as only three items in two categories remained. It is interesting to note that these items had a global quality. In other words, the items seemed to summarize the rating of

faculty teaching performance very succinctly from a student's perspective regarding what might be important (e.g., "Would you take another course with this instructor?").

A reduction of the consistency criterion to the 80% survivability level resulted in a longer, yet statistically consistent list of items. In other words, when applying a binomial probability test, assuming that each item has a 42% chance (the result of calculating the empirical probability for the present data) of being "kept" or "discarded", the result is significant, $p < .05$. The most apparent feature of the EC form was that its semantic content was almost identical to the UV questionnaire with regard to items pertaining to the rating of teaching performance (although the university form contained items not duplicated by the Echo-consistent form). Possible reasons for this result are (a) the students' criteria for what represents good teaching are basically identical to those of the faculty and administration, or (b) due to the fact that these students (being third and fourth year) have been exposed to several "standard" rating forms each school term. Consequently they may lack any former experience in assessing what constitutes good teaching since the students have simply internalized the criteria represented on the rating forms. Either alternative seems to be a likely explanation for the similarity between the Departmental and Echo-consistent forms.

One of the most interesting and significant results that came out of the present research was that the Echo-important form was judged to be preferred

over the Departmental and Echo-consistent forms. This preference seems to support the argument that it is inappropriate to employ normative methods in the evaluation of individual teaching performance. Although the items on the Echo-important form were chosen less consistently than the items on the Echo-consistent form, they were judged as the "most important" for the rating of teaching performance. The results indicate that when the statistical consistency criterion is reduced, the differences between the two forms are not only in their phrasing but also in their meaning. One difference between the Echo-consistent and Echo-important forms is that the Echo-consistent form contains some "global" or summary questions, whereas the items on the Echo-important form are of a more specific, descriptive nature.

Student-ratings may be reasonably reliable and valid for use by instructors for course improvement or by students choosing courses, but there is very little to support the use of student ratings of instruction for personnel decisions. This concern regarding the use of student ratings for administrative purposes is supported by Scriven (1988):

> Other concerns about validity remain, for example: (i) the general concern that student rating forms ask many questions about matters that students do not appear to be in any position to judge reliably; (ii) the fact that the overall rating of teaching merit by students is only statistically related to learning gains, a concern if one believes that it is inappropriate to use statistical indicators in personnel decisions; or (iii) the concern that the validation studies used to justify appeal to student ratings use questionable indicators instead of the true criterion. (For example, some of them correlate the student ratings with peer ratings of the merit of teachers instead of with superior learning gains.) (p. 5).

However, there is no evidence why student ratings should not be used in a formative manner. Through the open collaboration of students, faculty, and administration, considerable improvements can be made to the quality of courses and instruction (Abbott et al., 1990; L'Hommedieu et al., 1990). The most recent research substantiates this notion and emphasizes that a more effective formative evaluation product can be obtained through the use of more interactive techniques involving: feedback to students about rating results, facilitated group discussions, and/or interview methods (Abbott et al., 1990; L'Hommedieu et al., 1990).

Conclusion

Generally, the results of the present research indicate that the revised Echo technique for the development of student-generated teaching-performance rating questionnaires is efficient, fulfilling the conditions of reliability of production (through repeatability), and of content validity (through the results of the rating procedure). This new Echo method was able to elicit students' perceptions on the rating of instruction and, although the results were certainly not conclusive, they showed that students' perspectives of what constitutes an appropriate rating questionnaire for the evaluation of teaching performance can differ from the views of administrators and faculty. This research also showed that the evaluation of teaching performance is an extremely complex issue which would be better served by employing a standard method for the development of rating questionnaires rather than the use of standard questionnaires.

Any evaluation project should therefore ideally involve:

(a) a recognition of the contextual embeddedness of what constitutes "good teaching",

(b) all the stakeholders in an interactive methodological framework and,

(c) a variety of diverse evaluation methods and measures.

The results of such an evaluation may not always be generalizable, yet generalizability may not be a desirable goal when the evaluation product involves the assessment of individual performance.

REFERENCES

Abbott, R. D., Wulff, D. H., Nyquist, J. D., Ropp, V. A., & Hess, C. W. (1990). Satisfaction with processes of collecting student opinions about instruction: The student perspective. Journal of Educational Psychology, 82(2), 201-206.

Abrami, P. C., d'Apollonia, S., & Cohen, P. A. (1990). Validity of student ratings of instruction: What we know and what we do not. Journal of Educational Psychology, 82(2), 219-231.

Aleamoni, L. M. (1978). Development and factorial validation of the Arizona course/instructor evaluation questionnaire. Educational and Psychological Measurement, 38, 1063-1067.

Aleamoni, L. M. (1987). Student rating myths versus research facts. Journal of Personnel Evaluation in Education, 1, 111-119.

Aleamoni, L. M. & Hexner, P. Z. (1980). A review of the research on student evaluation and a report on the effect of different sets of instructions on student course and instructor evaluation. Instructional Science, 9, 67-84.

Arreola, R. A. (1987). The role of student government in faculty evaluation. In L. M. Aleamoni (Ed.). Techniques for Evaluating and Improving Instruction. New Directions for Teaching and Learning, no. 31. San Francisco: Jossey-Bass.

Aubrecht, J. D. (1979). Are student ratings of teacher effectiveness valid? IDEA Paper No. 2. Center for Faculty Evaluation and Development, Kansas State University.

Aubrecht, J. D. (1981). Reliability, validity and generalizability of student ratings of instruction. IDEA Paper No. 6. Center for Faculty Evaluation and Development, Kansas State University.

Barthol, R. P. and Bridge, R. G. (1968). The echo multi-response method for surveying value and influence patterns in groups. Psychological Reports, 22, 1345-1354.

Bavelas, A. (1942). A method for investigating individual and group ideology. Sociometry, 5, 371-377.

Bavelas, J B., Bavelas, A., and Schaefer, B. A. (1978). A Method for Constructing Student-Generated Faculty-Evaluation Questionnaires. Unpublished manuscript. The University of Victoria.

Bem, D. J., & Allen, A. (1974). On predicting some of the people some of the time: The search for cross-situational consistencies in behavior. Psychological Review, 81(6), 506-520.

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

Bendig, A. W. (1954). A factor analysis of student ratings of psychology instructors on the Purdue scale. The Journal of Educational Psychology, 45(7), 385-393.

Biddle, B. J. (1964). The integration of teacher effectiveness research. In B. J. Biddle, & W. J. Ellena (Eds.), Contemporary Research on Teacher Effectiveness (pp. 1-40). New York: Holt, Rinehart and Winston.

Blum, M. L. (1936). An investigation of the relation existing between students' grades and their ratings of the instructor's ability to teach. The Journal of Educational Psychology, 27, 217-221.

Boice, R. (1990-1991). Countering common misbeliefs about student evaluations of teaching. Teaching Excellence, 2(2), 1-2.

Cashin, W. E. (1988). Student ratings of teaching: A summary of the research. IDEA Paper No. 20. Center for Faculty Evaluation and Development, Kansas State University.

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

Costin, F., Greenough, W. T., & Menges, R. J. (1971). Student ratings of college teaching: Reliability,validity, and usefulness. Review of Educational Research, 41(5), 511-535.

Cranton, P. & Smith, R. A. (1990). Reconsidering the unit of analysis: A model of student ratings of instruction. Journal of Educational Psychology, 82(2), 207-212.

Dalhousie University Senate Committee on Instructional Development. (1991). Student ratings of instruction: Principles for practice. Halifax, N.S.: Dalhousie University, Office of Instructional Development and Technology.

Dewey, J. (1931). The Way Out of Educational Confusion. Cambridge: Harvard University Press.

Doyle, K. R. Jr. (1983). Evaluating Teaching. Toronto: D. C. Heath and Company.

English, H. B. & English, A. C. (1958). A Comprehensive Dictionary of Psychological and Psychoanalytical Terms. New York: Longmans, Green and Co.

Gitlin, A. & Smyth, J. (1989). Teacher Evaluation: Educative Alternatives. New York: The Falmer Press.

Guba, E. G. & Lincoln, Y. S. (1989). Fourth Generation Evaluation. Newbury Park CA: Sage.

Kerlinger, F. N. (1973). Foundations of Behavioral Research. (2nd ed.). New York: Holt, Rinehart and Winston, Inc.

L'Hommedieu, R., Menges, R. J., & Brinko, K. T. (1990). Methodological explanations for the modest effects of feedback from student ratings. Journal of Educational Psychology, 82(2), 232-241.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. Journal of Educational Psychology, 76(5), 707-754.

Marsh, H. W., Overall, J. U., & Kesler, S. P. (1979). Validity of student evaluations of instructional effectiveness: A comparison of faculty self-evaluations and evaluations by their students. Journal of Educational Psychology, 71(2), 149-160.

McKeachie, W. J. (1979). Student ratings of faculty: A reprise. Academe, 65, 384-397.

Moses, L. E. and Oakford, R. V. (1978). Tables of Random Permutations. Stanford: Stanford University 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(2), 250-261.

Perry, R. P. (1990). Special section: Instruction in higher education, introduction to the special section. Journal of Educational Psychology, 82(2), 183-188.

Remmers, H. H. & Elliott, D. N. (1949). The indiana college and university staff-evaluation program. School and Society, 70(1812), 168-171.

Schaefer, B. A., Bavelas, J. B., and Bavelas, A. (1980). Using Echo Technique to Construct Student-Generated Faculty-Evaluation Questionnaires. Teaching of Psychology, 7(2), 83-86.

Scriven, M. (1988). The validity of student ratings. Instructional Evaluation, 9(2), 5-18.

Scriven, M. (1991). Evaluation Thesaurus. (4th ed.). Newbury Park, CA: Sage.

Sidman, M. (1960). Tactics of Scientific Research. New York: Basic Books Inc.

Task Force on Teaching. (1992). Report of the task force on teaching. Victoria, B.C.: University of Victoria.

Whitehead, A. N. (1932/1955). The Aims of Education and Other Essays. London: Williams and Norgate Limited.

Wright, W. A. and O'Neil M. C. (1992). Improving summative student ratings of instruction practices. Journal of Staff Program and Organizational Development, 10(2), 75-85.

Appendix A

INSTRUCTIONS FOR ECHO SURVEY PHASE

A) The aim of this study is to develop a questionnaire for evaluating teaching performance. It should reflect the concerns of the students.

B) You should write down as many important questions as you can for the evaluation of professors who teach third and fourth year psychology classes. (i.e., questions that you wish were on course evaluation forms that allow you to say things (good or bad) which are important to you).

C) You are free to use any wording which you are comfortable with.

D) Please include response formats (i.e., the method in which you would like the question to be answered) for each question, or indicate which type of format you would prefer. The response format does not have to be a formal scale, be creative!

E) You are encouraged to go by your intuition when creating the questions, rather than attempting to be particularly "academic".

F) Write the questions the way they make the most sense to you and students like you.

G) Write as few or as many questions as you wish (3-15 would be appropriate).

H) Your participation is anonymous and no faculty will be involved in the study until after its conclusion, at which point the questions selected will not be traceable to you, their creators.

I) Please work on your own and return your responses anonymously to the psychology department under the name of "Echo study", or to me by appointment.

Appendix B

Final Raw Data

(136 items used in sorts 3-12 inclusive)

QUALITIES OF THE PROF'S TEACHING STYLE THAT EITHER ENHANCED OR DETRACTED FROM THE COURSE MATERIAL.
RESPONSE FORMAT: COMMENTS

IF I COULD CHANGE ANYTHING ABOUT THE COURSE I WOULD TAKE IT WITH ANOTHER PROFESSOR (IF AVAILABLE).
RESPONSE FORMAT: M/C - YES, PROBABLY, PROBABLY NOT, NO

RATE THE INSTRUCTOR COMPARED TO ALL YOUR OTHER INSTRUCTORS (THIS ONE IN RELATION TO THEM).
RESPONSE FORMAT: 1 (POOR) TO 10 (THE BEST)

WAS THIS TEACHER RECEPTIVE TO CLASS IDEAS?
RESPONSE FORMAT: M/C

DID THE PROFESSOR ACKNOWLEDGE THE FACT THAT YOU ARE ENROLED IN OTHER COURSES ASIDE FROM HIS/HERS WHEN HE/SHE CONSTRUCTED THE COURSE OUTLINE REQUIREMENTS? (I.E., WERE THE NUMBER OF ASSIGNMENTS AND EXAMS REASONABLE?)
RESPONSE FORMAT: CIRCLE: YES , NO , DOESN'T APPLY

WAS THE PROFESSOR ALWAYS AVAILABLE DURING HIS/HER OFFICE HOURS?
RESPONSE FORMAT: CIRCLE: ALWAYS , USUALLY , NEVER , DOESN'T APPLY

DID THE PROFESSOR ENCOURAGE STUDENTS TO ASK QUESTIONS AND PROVIDE COMMENTS?
RESPONSE FORMAT: CIRCLE: YES , NO , DOESN'T APPLY

IS THE PROF. APPROACHABLE REGARDING QUESTIONABLE CORRECT ANSWERS?
RESPONSE FORMAT: COMMENTS

WAS THE PROFESSOR APPROACHABLE?
RESPONSE FORMAT: COMMENTS

WILL THE PROF. MAKE TIME TO SEE YOU?
RESPONSE FORMAT: COMMENTS

RATE THE PROFESSOR ON HIS/HER CLARITY OF CONCEPTS. (I.E., CAN THEY COMMUNICATE INFORMATION EFFECTIVELY?)
RESPONSE FORMAT: M/C AND COMMENTS

RATE THE PROFESSOR ON HIS/HER WILLINGNESS TO ANSWER QUESTIONS.
RESPONSE FORMAT: M/C AND COMMENTS

RATE THE PROFESSOR ON HIS/HER AVAILABILITY.
RESPONSE FORMAT: M/C AND COMMENTS

RATE THE PROFESSOR ON HIS/HER ENCOURAGEMENT OF ASKING QUESTIONS.
RESPONSE FORMAT: M/C AND COMMENTS

WHAT IS YOUR PERSONAL IMPRESSION/OPINION OF THE PROF.?
RESPONSE FORMAT: COMMENT

WHAT DID YOU LIKE MOST ABOUT THE PROF.?
RESPONSE FORMAT: COMMENT

WHAT DID YOU LIKE LEAST ABOUT THE PROF.?
RESPONSE FORMAT: COMMENT

WOULD YOU TAKE ANOTHER COURSE FROM THIS PROFESSOR?
RESPONSE FORMAT: Y/N

DOES THE PROFESSOR PRESENT HIS/HER LECTURE IN AN ORGANIZED FASHION?
RESPONSE FORMAT: SHORT ANSWER

THE PROFESSOR'S APPROACH IN CREATING INTEREST IN THE SUBJECT MATTER WAS...
RESPONSE FORMAT: SHORT ANSWER

DOES THE INSTRUCTOR SEEM GENUINELY INTERESTED IN TEACHING THIS COURSE?
RESPONSE FORMAT: COMMENT

DOES THE INSTRUCTOR SEEM APPROACHABLE AND WILLING TO ANSWER QUESTIONS POSED BY STUDENTS?
RESPONSE FORMAT: COMMENT

DO YOU FEEL THAT THE INSTRUCTOR IS WELL SUITED TO TEACH THIS PARTICULAR COURSE?
RESPONSE FORMAT: COMMENT

HOW EFFECTIVE DO YOU FEEL THIS INSTRUCTOR'S TEACHING METHODS ARE?
RESPONSE FORMAT: COMMENT

WHAT DO YOU THINK THE INSTRUCTOR COULD DO TO IMPROVE THEIR TEACHING STYLE?
RESPONSE FORMAT: COMMENT

WHAT DO YOU THINK ARE THIS INSTRUCTOR'S:
RESPONSE FORMAT: STRENGTHS WEAKNESSES

THE INSTRUCTOR IS ABLE TO EXPLAIN A CONCEPT OR IDEA FROM SEVERAL PERSPECTIVES AND TO EXPAND UPON A CONCEPT OR IDEA IN ORDER TO CLARIFY IT.
RESPONSE FORMAT: SCALE OF 1-5

THE INSTRUCTOR APPEARS TO HAVE A GOOD UNDERSTANDING OF THE MATERIAL BEING DISCUSSED/TAUGHT AND DOES NOT SUBSTITUTE NAMING THINGS FOR EXPLAINING THINGS.
RESPONSE FORMAT: SCALE OF 1-5

THE INSTRUCTOR IS ABLE TO THINK LOGICALLY AND COHERENTLY ABOUT THE MATERIAL BEING PRESENTED AND IS ABLE TO SHARE THIS THOUGHT PROCESS WITH THE CLASS.
RESPONSE FORMAT: SCALE OF 1-5

DOES THE TEACHER EFFECTIVELY CONVEY HIS OR HER KNOWLEDGE OF THE COURSE MATERIAL?
RESPONSE FORMAT: A) VERY WELL, B) WELL, C) POORLY, D) VERY POORLY

HOW COULD THE TEACHER IMPROVE UPON THE EFFECTIVENESS OF HIS OR HER PRESENTATION OF THE MATERIAL?
RESPONSE FORMAT: COMMENT

DID THE INSTRUCTOR HAVE A GOOD RAPPORT WITH THE CLASS?
RESPONSE FORMAT: YES/NO - IF NO, EXPAND/SUGGESTIONS.

WAS THE INSTRUCTOR APPROACHABLE?
RESPONSE FORMAT: YES/NO - IF NO, EXPAND/SUGGESTIONS.

IS THE PROFESSOR'S LECTURE TECHNIQUE INTERESTING?
RESPONSE FORMAT: A) EXCELLENT, B) VERY GOOD, C) GOOD, D) FAIR, E) POOR, F) VERY POOR. ADDITIONAL COMMENTS.

IS THE PROFESSOR EASILY ACCESSIBLE?
RESPONSE FORMAT: A) EXCELLENT, B) VERY GOOD, C) GOOD, D) FAIR, E) POOR, F) VERY POOR. ADDITIONAL COMMENTS.

IS THE PROFESSOR EASY TO TALK TO IN CLASS?
RESPONSE FORMAT: A) EXCELLENT, B) VERY GOOD, C) GOOD, D) FAIR, E) POOR, F) VERY POOR. ADDITIONAL COMMENTS.

THE INSTRUCTOR'S PREPARATION FOR EACH DAY'S LECTURE WAS...?
RESPONSE FORMAT: A) EXCELLENT, B) VERY GOOD, C) GOOD, D) FAIR, E) POOR, F) VERY POOR. ADDITIONAL COMMENTS.

COMPARED TO OTHER INSTRUCTORS, HOW WOULD YOU RATE YOUR INSTRUCTOR IN TERMS OF GENERAL, OVERALL TEACHING EFFECTIVENESS?
RESPONSE FORMAT: A) EXCELLENT, B) VERY GOOD, C) GOOD, D) FAIR, E) POOR, F) VERY POOR. ADDITIONAL COMMENTS.

WHAT WERE THE STRONG POINTS - WEAK POINTS OF THE INSTRUCTOR?
RESPONSE FORMAT: COMMENTS.

DID THE INSTRUCTOR ANSWER QUESTIONS CLEARLY?
RESPONSE FORMAT: COMMENTS.

HOW WOULD YOU RATE THE LEVEL OF PRESENTATION BY THE INSTRUCTOR?
RESPONSE FORMAT: COMMENTS.

WAS THE PROF. OPEN TO QUESTIONS WITHIN THE CLASS PERIOD?
RESPONSE FORMAT: SCALE/COMMENTS

DID THE PROF. SEEM WELL-ACQUAINTED WITH THE MATERIAL THAT HE/SHE WAS PRESENTING?
RESPONSE FORMAT: SCALE/COMMENTS

WAS THE PROF. KNOWLEDGEABLE WITH THE TOPICS PRESENTED, OR WITH RELATED TOPICS ENOUGH TO ADDRESS REASONABLE QUESTIONS?
RESPONSE FORMAT: SCALE/COMMENTS

DID THE PROF. HAVE SUFFICIENT OFFICE HOURS FOR THE DIFFICULTY OF THE COURSE OR FOR THE SIZE OF ENROLMENT?
RESPONSE FORMAT: SCALE/COMMENTS

DID THE PROF. OFFER OTHER SOURCES FOR REFERENCE TO RELATED TOPICS OF INTEREST?
RESPONSE FORMAT: SCALE/COMMENTS

DID THE PROF. LEAVE AMPLE TIME IN THE OVERALL TIMING OF THE CLASS FOR POSSIBLE QUESTIONS OR PROBLEMS?
RESPONSE FORMAT: SCALE/COMMENTS

DID THE PROF. APPEAR TO HAVE AN INTEREST IN THE TOPICS PRESENTED AND TO THE SUBJECT MATTER OF THE COURSE IN GENERAL?
RESPONSE FORMAT: SCALE/COMMENTS

HOW CLOSELY DID THE PROFESSOR STICK TO WHAT WAS WRITTEN IN THE TEXT?
RESPONSE FORMAT: SCALE OF 1-10: 1 = DID NOT FOLLOW THE TEXT AT ALL, 10 = PROFESSOR FOLLOWED THE TEXT EXACTLY

IF YOU CONSULTED WITH THE PROFESSOR, HOW MUCH BETTER DID YOU UNDERSTAND THE MATERIAL AFTER CONSULTING THE PROFESSOR?
RESPONSE FORMAT: SCALE OF 1-10: 1 = UNDERSTOOD SAME AMOUNT AS BEFORE, 10 = UNDERSTOOD MATERIAL PERFECTLY

DID YOU ENJOY THE COURSE AND IF SO OR IF NOT, WAS THE INSTRUCTOR A CONTRIBUTING FACTOR?
RESPONSE FORMAT: WRITTEN ANSWER

WAS THE INSTRUCTOR EASY TO APPROACH, SPEAK WITH AND ASK FOR HELP?
RESPONSE FORMAT: A) VERY GOOD, B) GOOD, C) MODERATE, D) POOR, E) VERY POOR

DID THE INSTRUCTOR MAKE THE INFORMATION AS INTERESTING AS POSSIBLE? WHAT CHANGES COULD BE MADE?
RESPONSE FORMAT: SCALE OF 1-5 AND WRITTEN COMMENT

DID YOU THINK THAT THE INSTRUCTOR WAS AN EFFECTIVE TEACHER?
RESPONSE FORMAT: YES/NO - AND WHY (WRITTEN)

WAS THE VOICE OF THE PROFESSOR;
RESPONSE FORMAT: DIFFICULT.....CLEAR.....GOOD - TO UNDERSTAND

AT WHAT PACE DID THE PROFESSOR TEACH?
RESPONSE FORMAT: SLOW...AVERAGE...FAST

DO YOU FEEL THE PROFESSOR PRESENTED THE LECTURES ON COURSE MATERIAL EFFICIENTLY?
RESPONSE FORMAT: SHORT ANSWER

WAS THE PROFESSOR READILY AVAILABLE AND WILLING TO GIVE EXTRA HELP OUTSIDE OF CLASS?
RESPONSE FORMAT: SHORT ANSWER

DOES THE PROFESSOR PRESENT THE COURSE MATERIAL IN AN INTERESTING MANNER?
RESPONSE FORMAT: SHORT ANSWER

DOES THE PROFESSOR SHOW INTEREST IN THE COURSE MATERIAL?
RESPONSE FORMAT: SHORT ANSWER

DO YOU FEEL THAT THIS PROFESSOR HAS SUFFICIENTLY BROKEN DOWN THE MATERIAL OF THE COURSE IN A MANNER THAT RENDERS IT UNDERSTANDABLE?
RESPONSE FORMAT: YES/NO - COMMENT

WAS THE PROFESSOR APPROACHABLE WHEN/IF YOU NEEDED HELP WITH THE MATERIAL?
RESPONSE FORMAT: YES/NO - COMMENT

DID THE PROFESSOR REVIEW AND OUTLINE WHAT WAS EXPECTED OF YOU FOR EACH EXAM?
RESPONSE FORMAT: YES/NO - COMMENT

WHY DID YOU CHOOSE THIS PROFESSOR AS OPPOSED TO ANY OTHER ONE?
RESPONSE FORMAT: A) I HAVE HAD HIM/HER BEFORE, B) THERE WAS ONLY ONE SECTION AVAILABLE, C) SCHEDULE CONFLICT, D) WORD OF MOUTH, D) OTHER (PLEASE SPECIFY)_____.

PLEASE RATE YOUR PROFESSOR ON APPROACHABILITY (I.E., CAN YOU APPROACH HIM/HER WITH QUESTIONS).
RESPONSE FORMAT: SCALE FROM 1-3 (WITH THREE BEING THE HIGHEST).

PLEASE RATE YOUR PROFESSOR ON PATIENCE.
RESPONSE FORMAT: SCALE FROM 1-3 (WITH THREE BEING THE HIGHEST).

PLEASE RATE YOUR PROFESSOR ON AVAILABILITY (OTHER THAN CLASS HOURS).
RESPONSE FORMAT: SCALE FROM 1-3 (WITH THREE BEING THE HIGHEST).

PLEASE RATE YOUR PROFESSOR ON SPEED OF GRADING AND RETURNING WORK.
RESPONSE FORMAT: SCALE FROM 1-3 (WITH THREE BEING THE HIGHEST).

IS THIS PROFESSOR EASY TO UNDERSTAND (WAY THEY SPEAK)?
RESPONSE FORMAT: YES/NO

IS THIS PROFESSOR USUALLY ON TIME TO CLASS?
RESPONSE FORMAT: YES/NO

IS THIS PROFESSOR AVAILABLE OUTSIDE OF HIS/HER OFFICE HOURS?
RESPONSE FORMAT: YES/NO

IS THIS PROFESSOR QUICK IN GRADING ASSIGNMENTS?
RESPONSE FORMAT: YES/NO

DID YOU THINK YOUR PROFESSOR'S GRADING POLICY WAS FAIR? IF NOT, PLEASE STATE REASONS.
RESPONSE FORMAT: SHORT ANSWER

DID YOUR PROFESSOR'S LECTURES CLOSELY FOLLOW THE ASSIGNED READINGS, AND DID THIS MAKE UNDERSTANDING OF THE READINGS EASIER OR MORE DIFFICULT?
RESPONSE FORMAT: SHORT ANSWER OR YES/NO AND MULTIPLE CHOICE.

DID YOUR PROFESSOR SPEAK IN A CLEARLY AUDIBLE VOICE (ESPECIALLY IF YOU SAT IN THE BACK OF THE ROOM)?
RESPONSE FORMAT: SHORT ANSWER OR YES/NO AND MULTIPLE CHOICE.

DO YOU THINK YOUR PROFESSOR'S EFFECTIVENESS AS AN INSTRUCTOR INFLUENCED YOUR GRADE IN THIS COURSE? IF SO, WAS THE INFLUENCE POSITIVE OR NEGATIVE?
RESPONSE FORMAT: SHORT ANSWER

WERE YOU SATISFIED WITH HOW YOUR PROFESSOR ANSWERED QUESTIONS IN CLASS? IF NOT, PLEASE STATE REASONS.
RESPONSE FORMAT: SHORT ANSWER

WHAT WAS THE BEST THING ABOUT YOUR PROFESSOR?
RESPONSE FORMAT: SHORT ANSWER

WHAT ASPECT(S) OF YOUR PROFESSOR NEED(S) THE MOST IMPROVEMENT?
RESPONSE FORMAT: SHORT ANSWER

HOW WELL DID THE INSTRUCTOR SPEAK?
RESPONSE FORMAT:
5_____1
CLEARLY.....UNCLEARLY

HOW PREPARED WAS THIS INSTRUCTOR?
RESPONSE FORMAT:
5_____1
ALWAYS.....NEVER

HOW FLEXIBLE WAS THE INSTRUCTOR TO THE CONTENT OF THIS COURSE REGARDING QUESTION, ANSWER, DISCUSSION TIME, AND WHY?
RESPONSE FORMAT: ___EXCELLENT ___VERY GOOD
___APPROPRIATE ___LACKING ___POOR, PLUS SHORT ANSWER.

HOW FLEXIBLE WAS THE INSTRUCTOR TO THE CONTENT OF THIS COURSE REGARDING COORDINATION AND COURSE OUTLINE, AND WHY?
RESPONSE FORMAT: ___EXCELLENT ___VERY GOOD
___APPROPRIATE ___LACKING ___POOR, PLUS SHORT ANSWER.

HOW FLEXIBLE WAS THE INSTRUCTOR TO THE CONTENT OF THIS COURSE REGARDING PACE OF LECTURES (I.E., DID THE LECTURER FALL BEHIND), AND WHY?
RESPONSE FORMAT: ___EXCELLENT ___VERY GOOD
___APPROPRIATE ___LACKING ___POOR, PLUS SHORT ANSWER.

WAS THE INSTRUCTOR'S KNOWLEDGE OF MATERIAL EFFECTIVE IN TEACHING COURSE, WHY?
RESPONSE FORMAT: ___EXCELLENT ___VERY GOOD
___APPROPRIATE ___LACKING ___POOR, PLUS SHORT ANSWER.

HOW APPROACHABLE WAS THE INSTRUCTOR?
RESPONSE FORMAT:
5_____1
ALWAYS.....NEVER

WERE RESPONSES TO STUDENTS' QUESTIONS WELL STRUCTURED BY THE INSTRUCTOR?
RESPONSE FORMAT:
5_____1
ALWAYS.....NEVER

DID THE INSTRUCTOR LECTURE ON TOPIC?
RESPONSE FORMAT:
5_____1
ALWAYS.....NEVER

DID THE INSTRUCTOR GENERATE CLASS DISCUSSIONS?
RESPONSE FORMAT:
OFTEN....SOMETIMES....OCCASIONALLY....NEVER

WOULD YOU TAKE ANOTHER COURSE WITH THIS INSTRUCTOR?
RESPONSE FORMAT: YES/NO

HOW INTERESTING WAS THIS INSTRUCTOR'S TEACHING STYLE? (IF RATED AT 5 OR BELOW, HOW MIGHT THEY IMPROVE?)
RESPONSE FORMAT: SCALE 1-10, COMMENT

DID YOU FEEL THE INSTRUCTOR WAS CONDESCENDING?
RESPONSE FORMAT: CIRCLE ONE:
NEVER..HARDLY EVER..SOMETIMES..OFTEN..ALWAYS

DID YOU FEEL THAT THE INSTRUCTOR WAS APPROACHABLE?
RESPONSE FORMAT: CIRCLE ONE,HIGHLY..MODERATELY..MILDLY ..NOT AT ALL..DON'T KNOW

HOW MANY TIMES DID YOU SEE THE INSTRUCTOR OUTSIDE OF CLASS?
RESPONSE FORMAT: NUMBER OF TIMES___

IF YOU WERE GIVEN THE CHOICE WOULD YOU HIRE THIS INSTRUCTOR TO TEACH YOU?
RESPONSE FORMAT: Y/N - COMMENT

IN THE FUTURE WOULD YOU TAKE A COURSE TAUGHT BY THE SAME INSTRUCTOR?
RESPONSE FORMAT: Y/N - COMMENT

WAS THE PROFESSOR'S METHOD OF PRESENTATION EFFECTIVE?
RESPONSE FORMAT: Y/N - COMMENT

DID THE PROFESSOR MAKE THE MATERIAL TO BE TAUGHT AS INTERESTING AS POSSIBLE?
RESPONSE FORMAT: VERY INTERESTING...SOMEWHAT INTERESTING ...NOT INTERESTING

DID THE PROFESSOR SCHEDULE ADEQUATE OFFICE HOURS SO AS TO ALLOW ALL STUDENTS THE OPPORTUNITY TO CONSULT?
RESPONSE FORMAT: Y/N

DID THE PROFESSOR USE THE REQUIRED TEXTBOOK AS MUCH AS HE/SHE COULD WITHOUT BEING TOO REPETITIVE?
RESPONSE FORMAT: YES/NO/ADEQUATE

DID THE PROFESSOR TEACH A LOT OF MATERIAL EXTRANEOUS TO EXAM REQUIREMENTS?
RESPONSE FORMAT: YES/NO/A FAIR AMOUNT

DID THE PROFESSOR CLEARLY EXPLAIN EXAMINATION FORMATS?
RESPONSE FORMAT: YES/NO/ADEQUATE

WAS THE INSTRUCTOR GENERALLY RESPECTFUL OF STUDENTS AS INTELLIGENT INDIVIDUALS (I.E., RESPOND TO QUESTIONS RAISED BY STUDENTS, ENCOURAGED STUDENTS' OPINIONS)?
RESPONSE FORMAT: A) ALWAYS, B) USUALLY, C) SELDOM, D) NEVER

DOES THE INSTRUCTOR STAY ON TOPIC AS DEFINED BY THE COURSE OUTLINE?
RESPONSE FORMAT: A) ALWAYS, B) USUALLY, C) SELDOM, D) NEVER

DOES THE PROFESSOR TREAT HIS/HER STUDENTS WITH RESPECT (I.E., DOESN'T TALK DOWN AND PATRONIZE, DOES LISTEN TO THEIR IDEAS ETC.)?
RESPONSE FORMAT: SHORT ANSWER

DOES THE PROFESSOR MAKE HIM/HERSELF AVAILABLE TO ANSWER ANY QUESTIONS OR HELP WITH MISUNDERSTANDINGS?
RESPONSE FORMAT: SHORT ANSWER

DOES THE PROFESSOR DO HIS/HER BEST TO MAKE THE COURSE INTERESTING?
RESPONSE FORMAT: SHORT ANSWER

IS THE PROFESSOR ENTHUSIASTIC ABOUT THE SUBJECT MATTER?
RESPONSE FORMAT: SHORT ANSWER

DOES THE PROFESSOR COVER MATERIAL AT A REASONABLE PACE?
RESPONSE FORMAT: SHORT ANSWER

IS THE PROFESSOR CLEAR ABOUT WHAT WILL BE EXAMINABLE AND EXPLAIN WHAT THE EXAMS WILL BE LIKE?
RESPONSE FORMAT: SHORT ANSWER

DOES THE WAY IN WHICH YOUR INSTRUCTOR DRESSES AND PRESENTS HIM/HERSELF AFFECT YOUR ATTITUDE TOWARD YOUR INSTRUCTOR?
RESPONSE FORMAT: ANSWER WITH DIRECT COMMENT

DO YOU HAVE RESPECT FOR YOUR TEACHER'S INSTRUCTION METHODS AND FOR HIM/HER AS A WHOLE?
RESPONSE FORMAT: ANSWER WITH DIRECT COMMENT

IN YOUR OPINION SHOULD THIS PROFESSOR TEACH THIS COURSE AGAIN?
RESPONSE FORMAT: WRITE ANSWER IN YOUR OWN WORDS

DOES THIS PROFESSOR'S TEACHING SKILLS WARRANT A POSITION AT THE UNIVERSITY LEVEL - WHY, WHY NOT?
RESPONSE FORMAT:WRITE ANSWER IN YOUR OWN WORDS

DID THE PROFESSOR'S STYLE DISCOURAGE OR ENCOURAGE YOUR ATTENDANCE IN ANY SPECIFIC WAY - HOW?
RESPONSE FORMAT: WRITE ANSWER IN YOUR OWN WORDS

DOES THE INSTRUCTOR INSTILL CONFIDENCE IN YOU OF HIS/HER TEACHING ABILITY?
RESPONSE FORMAT: WRITE ANSWER IN YOUR OWN WORDS

DID THE PROFESSOR ADHERE TO THE LECTURE MATERIAL OR DID HE/SHE GO OFF ON IRRELEVANT TANGENTS?
RESPONSE FORMAT: Y/N - COMMENT

DID THE PROFESSOR COME TO BEGIN CLASS ON TIME AND FINISH ON TIME (I.E., NOT GO TOO LONG OR FINISH TOO EARLY)?
RESPONSE FORMAT: COMMENT

DID THE PROFESSOR SEEM COMFORTABLE/RESPONSIVE IN ACCEPTING QUESTIONS OR COMMENTS DURING THE LECTURES?
RESPONSE FORMAT: Y/N - COMMENT

DID THE PROFESSOR PROVIDE TIME IN LECTURES FOR QUESTIONS/DISCUSSIONS?
RESPONSE FORMAT: Y/N - COMMENT

HOW MUCH EFFORT DO YOU FEEL THE PROF. PUT INTO THIS COURSE?
RESPONSE FORMAT:
1.....10
LITTLE EFFORT_____MAXIMUM EFFORT

DOES THE PROFESSOR HAVE CONSIDERATION AND RESPECT FOR STUDENTS WHEN QUESTIONS ARE RAISED?
RESPONSE FORMAT: SCALE OF 1-5

DO YOU FEEL THE PROFESSOR IS BEING FAIR IN EVALUATION?
RESPONSE FORMAT: SCALE OF 1-5

DO THE PROFESSOR'S EXPECTATIONS EXCEED THE ABILITIES OF THE STUDENTS?
RESPONSE FORMAT: Y/N - COMMENT

DOES THE PROFESSOR PRESENT INFORMATION IN A MANNER WHICH IS EASY TO COMPREHEND?
RESPONSE FORMAT: Y/N - COMMENT

IS YOUR PROFESSOR APPROACHABLE (I.E., WHEN YOU HAVE A PROBLEM)?
RESPONSE FORMAT: Y/N - COMMENT

DOES THE TEACHER CONVEY MATERIAL CLEARLY?
RESPONSE FORMAT: COMMENT & SCALE
1.....5
NEVER_____ALWAYS

DOES THE TEACHER MAKE THE COURSE INTERESTING?
RESPONSE FORMAT: COMMENT & SCALE
1.....5
NEVER_____ALWAYS

IS THE PROFESSOR AVAILABLE FOR HELP?
RESPONSE FORMAT: COMMENT & SCALE
1.....5
NEVER_____ALWAYS

IS THE PROFESSOR AVAILABLE WHEN HE/SHE IS SUPPOSED TO HAVE OFFICE HOURS?
RESPONSE FORMAT: A) YES, B) NO, C)MOST OF THE TIME, D) SOMETIMES

IS THE PROFESSOR ABLE TO MAKE A LECTURE THAT FOLLOWS IN CHRONOLOGICAL ORDER?
RESPONSE FORMAT: A) YES, B) NO, C)MOST OF THE TIME, D) SOMETIMES

DOES THE INSTRUCTOR COVER MATERIAL AT A COMFORTABLE PACE, TO ASSURE UNDERSTANDING OF THE MATERIAL COVERED?
RESPONSE FORMAT: WRITTEN RESPONSE

IS THE INSTRUCTOR A COMPETENT SPEAKER?
RESPONSE FORMAT: WRITTEN RESPONSE

ARE THE INSTRUCTOR'S METHODS OF EVALUATION (EXAMS) FAIR (I.E., DOES IT REFLECT WHAT HAS BEEN TAUGHT IN LECTURES OR ASSIGNED IN READINGS)?
RESPONSE FORMAT: WRITTEN RESPONSE

DID THE TEACHER HAND BACK TESTS AND PROJECTS IN A REASONABLE AMOUNT OF TIME?
RESPONSE FORMAT: Y/N

WHAT IS THE CLARITY RATING OF INSTRUCTION GIVEN? (I.E., ARE YOU ABLE TO UNDERSTAND THE INSTRUCTOR'S LECTURES)
RESPONSE FORMAT: 1 (POOR) TO 10 (THE BEST)

Appendix C

REVISED "ECHO" SORTING INSTRUCTIONS

(Sorting groups 3-12 inclusive)

Background:

This research focuses on the student-rating of faculty teaching performance. Until now the psychology department has been using a questionnaire which was constructed by the faculty. The goal of my research is to compile a student-generated questionnaire for the evaluation of faculty teaching performance. This will give the students an opportunity to have some direct input into the generation of a faculty evaluation questionnaire which should therefore better reflect the students' concerns.

Your task is a sorting procedure comprised of four basic components. I will explain the entire procedure briefly and then will explain each step as you get to it. Here is a set of items generated by a representative sample of students registered in third and fourth year psychology classes. These items reflect those students' concerns regarding what they feel are important questions for the evaluation of faculty teaching performance. For practical purposes your focus here is strictly the student-rating of faculty teaching performance - not courses in general, or anything else.

The entire procedure will be a group effort and should take approximately one and one half hours. Your final goal is to have created an assortment of categories where each category contains items which are homogeneous (the items mean the same thing). You will then decide which question in each category is the most representative of that category. Finally, you will rank order the categories as to: which you feel is the most important and should appear on a faculty evaluation questionnaire - to which is the least important.

Instructions:

1. You each have a portion of the items in front of you. Please take turns reading out loud the item on each card. You may disregard the response format for the time being unless you need it in order to better understand the item. As each item is read, you should decide, as a group, whether or not the card should be kept or discarded. You should keep the item if it is: clear to understand and appeals to your intuitive judgement as a appropriate item for the rating of faculty teaching performance. You should discard the item if it is: poorly phrased, weird, ambiguous, hard to understand, voicing several concerns at once, or generally does not appeal to you as being an appropriate question for the rating of faculty

teaching performance. If at any time all of you do not agree on whether to keep or discard an item, simply put it in a miscellaneous pile for now and we will deal with those later.

(REMIND THE GROUP OF THEIR PRIMARY GOAL/FOCUS BETWEEN EACH STEP OF THE PROCEDURE)

2. Now you need to sort the remaining cards into categories. If necessary you may still discard items. You may have as many or as few categories as you require. You may find that some of the categories will contain more than one item - that is alright. Although your goal is to have each category homogeneous (all the items mean the same thing) do not worry about heterogeneity - you will be checking the categories later. I will read out the items to you and you simply need to tell me what category they should go into. (The miscellaneous category was always done last - the items were either categorized or discarded).

(REMIND THE GROUP OF THEIR PRIMARY GOAL/FOCUS BETWEEN EACH STEP OF THE PROCEDURE)

3. Please go through each category one at a time, beginning with the larger ones and, by taking the first two cards and comparing them, decide which is most representative of that category. Match consecutive items with whichever one you keep, until you have finished the category. If you feel at any time that a card should be in a different category, feel free to remove/re-categorize it.

(REMIND THE GROUP OF THEIR PRIMARY GOAL/FOCUS BETWEEN EACH STEP OF THE PROCEDURE)

4. Finally I ask that you rank order the categories (using only the representative question as your criterion) in the order of most to least important (i.e., those which definitely should appear on an evaluation questionnaire for the rating of faculty teaching performance, to those that do not need to).

Appendix D

STUDENT SIGN-UP SHEET FOR 'ECHO' STUDY

(Research Project No. 81-92)

NAME (Please Print)	PHONE #	ID.#

Appendix E

Echo-Consistent Questionnaire (EC)

QUESTIONNAIRE FOR STUDENT-RATING
OF FACULTY PERFORMANCE

Did the professor clearly explain examination formats?

_____

Does the professor show interest in the course material?

_____

Would you take another course with this instructor?

_____

Did the teacher hand back tests and projects in a reasonable amount of time?

_____

Was the professor available during his/her office hours?

_____

Does the instructor stay on topic as defined by the course outline?

_____

Did the professor encourage students to ask questions and provide comments?

_____

Does the professor present his/her lecture in an organized fashion?

_____

Did you think your professor's grading policy was fair?

Appendix F

Echo-Important Questionnaire (EI)

QUESTIONNAIRE FOR STUDENT-RATING
OF FACULTY PERFORMANCE

Is the professor clear about what will be examinable and explain what the exams will be like?

_____

Does the instructor seem genuinely interested in teaching this course?

_____

Does the professor present information in a manner which is easy to comprehend?

_____

Does the instructor cover material at a comfortable pace, to assure understanding of the material covered?

_____

Did the instructor answer questions clearly?

_____

Did the professor adhere to the lecture material or did he/she go off on irrelevant tangents?

_____

How prepared was this instructor?

_____

Does the professor present his/her lecture in an organized fashion?

_____

Are the instructor's methods of evaluation fair?

Appendix G

UNIVERSITY OF VICTORIA

DEPARTMENT OF PSYCHOLOGY

Teaching Effectiveness Questionnaire

This survey consists of a number of statements concerning various aspects of teaching effectiveness and is being used as part of the evaluation of instructors in the department of psychology. Please fill in the appropriate space on your computer answer sheet, making sure you also fill in the course number and section. Please do not fill in the spaces relating to your name or student number, so that your anonymity is guaranteed. Be careful in answering each statement as the result of this teaching evaluation could play a part in the decision made about your instructor and contract renewal, promotion and tenure. Thank you for being part of this survey.

*1. The instructor gave grades, with corrective comments, on all work submitted which counted towards the final grade.
   (a) Always
   (b) Usually
   (c) Seldom
   (d) Never
   (e) Don't know/does not apply to this course

*2. Given the size of the class, the instructor returned all student work submitted within a reasonable period of time.
   (a) Always
   (b) Usually
   (c) Seldom
   (d) Never
   (e) Don't know/does not apply to this course

*3. The instructor informed the students of grading criteria and procedures within the first 2 weeks of the course.
   (a) Gave full details
   (b) Gave general outline
   (c) Gave sketchy information
   (d) Gave no information
   (e) Don't know/does not apply to this course

*4. The instructor informed or made arrangements to inform the students of class cancellations.
(a) Always
(b) Usually
(c) Seldom
(d) Never
(e) Don't know/does not apply to this course

*5. The instructor was available for out-of-class consultation if desired by the student (either informally after class, scheduled office hours or by appointment).
(a) Always
(b) Usually
(c) Seldom
(d) Never
(e) Don't know/does not apply to this course

*Questions refer to the Senate recommendations for minimum teaching standards.

6. Generally, this instructor has made the course as interesting as the subject matter would allow:
(a) Strongly agree
(b) Agree
(c) Disagree
(d) Strongly disagree
(e) Don't know/does not apply

7. The instructor explained concepts and principles clearly.
(a) Strongly agree
(b) Agree
(c) Disagree
(d) Strongly disagree
(e) Don't know/does not apply

8. The instructor was well prepared for class.
(a) Strongly agree
(b) Agree
(c) Disagree
(d) Strongly disagree
(e) Don't know/does not apply

9. I would recommend this instructor to a friend whose interests were similar to mine.
   (a) Strongly agree
   (b) Agree
   (c) Disagree
   (d) Strongly disagree
   (e) Don't know/does not apply

10. Teaching methods were used which were appropriate to the content of this course.
   (a) Strongly agree
   (b) Agree
   (c) Disagree
   (d) Strongly disagree
   (e) Don't know/does not apply

NOTE: If you feel that you do not have enough information to answer the following two questions, please leave totally blank. Thank you.

11. Relative to other instructors in Psychology courses which I have taken, I would rate this instructor as:
   (a) Outstanding
   (b) Very good
   (c) Good
   (d) Fair
   (e) Poor

12. Relative to other instructors from other disciplines whose courses I have taken, I would rate this instructor as:
   (a) Outstanding
   (b) Very good
   (c) Good
   (d) Fair
   (e) Poor

Additional comments or extra questions responses: (use back of page if necessary)

Appendix H

Modified Psychology Department Questionnaire (MD)

QUESTIONNAIRE FOR STUDENT-RATING
OF FACULTY PERFORMANCE

Did the instructor inform the students of grading criteria and procedures within the first 2 weeks of the course?

_____

Generally, did this instructor make the course as interesting as the subject matter would allow?

_____

Was the instructor well prepared for class?

_____

Was the instructor available for out-of-class consultation if desired by the student (either informally after class, scheduled office hours or by appointment)?

_____

Did the instructor explain concepts and principles clearly?

_____

Would you recommend this instructor to a friend whose interests are similar to yours?

_____

Given the size of the class, did the instructor return all student work submitted within a reasonable amount of time?

_____

Were the teaching methods used appropriate to the content of this course?

_____

Did the instructor give grades, with corrective comments, on all work submitted which counted towards the final grade?

Appendix I

Verbal Instructions for Rating Phase

My name is Werner and I am a graduate student in psychology.

Dr. _____ has kindly given me this class time to conduct my research.

I would therefore very much appreciate your participation in my research which has to do with the evaluation of faculty teaching performance.

Your participation is of course voluntary and should only take a few minutes of your time.

Please carefully read and follow the instructions on the cover sheet of the material you receive.

If you would like to make a brief written comment, please do so on the cover sheet.

If you do not wish to participate simply return the material unmarked.

Whether or not you have participated, please return the material to me face down on your desks for collection - do not take the forms with you.

I'D LIKE TO THANK YOU IN ADVANCE FOR YOUR PARTICIPATION.

Appendix J

Rating Phase Subject
Response/Instructions Cover Sheet

THE FOLLOWING ITEMS HAVE BEEN GENERATED FOR THE CONSTRUCTION OF A STUDENT-RATING QUESTIONNAIRE, TO BE USED FOR THE EVALUATION OF YOUR INSTRUCTOR'S TEACHING PERFORMANCE.

INSTRUCTIONS - PLEASE READ CAREFULLY:

1. Read the attached questionnaires.

2. Indicate with an "x" in the appropriate space below, which of the two questionnaires better addresses your concerns as a student.

3. Please return all pages.

THANK YOU.

FORM A ___

FORM B ___

VITA

Surname: Müller-Clemm Given Names: Werner Johannes
Place of Birth: Vancouver, British Columbia, Canada
Date of Birth: 08 April 1960

Educational Institutions Attended:

Fresno City College (CA)	1988 to 1989
University of Victoria	1989 to 1991
University of Victoria	1991 to 1993

Degrees Awarded:

B.Sc.	University of Victoria	1991
M.A.	University of Victoria	1993

Honors and Awards:

Fresno City College Award for
Scholastic Excellence 1988

Fresno City College Award for
Scholastic Excellence 1989

Honorary Mention for paper presented
at the Annual Convention of the
Canadian Psychological Association 1993

Publications:

Conference Paper Presentations:

Müller-Clemm, W. J. & Tolman, C. W. (1992, June). Stumpf and Titchener on affection: Methodologies apart. Paper presented at the annual convention of the Canadian Psychological Association, Quebec City, PQ.

Müller-Clemm, W. J. (1993, May). Value pluralism in fourth generation evaluation. Paper presented at the annual convention of the Canadian Psychological Association, Montreal, PQ.

PARTIAL COPYRIGHT LICENSE

I hereby grant the right to lend my thesis to users of the University of Victoria Library, and to make single copies only for such users or in response to a request from the library of any other university, or similar institution, on its behalf or for one of its users. I further agree that permission for the extensive copying of this thesis for scholarly purposes may be granted by me or a member of the university designated by me. It is understood that the copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Tile of Thesis: A Student-generated Questionnaire for the Evaluation of Teaching in Psychology.

Author: [redacted]

Werner J. Müller-Clemm

02 September 1993