May 21, 1956

To the Graduate Council:

I am submitting herewith a thesis written by Emmoran Benjamin Cobb entitled "Construction of a Forced-Choice University Instructor Rating Scale." I recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Psychology.

Edward E. Cureton
Major Professor

We have read this thesis and recommend its acceptance:

J. M. Cotton Jr.
H. L. Parsons
W. O. Jenkins
Howard S. Emerson

Accepted for the Council:

E. G. Sexton
Dean of the Graduate School

CONSTRUCTION OF A FORCED-CHOICE UNIVERSITY INSTRUCTOR RATING SCALE

A THESIS

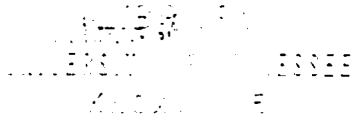
Submitted to
The Graduate Council
of
The University of Tennessee
in
Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

---

by

Emmoran Benjamin Cobb

August 1956



ACKNOWLEDGMENT

To my committee, Drs. E. E. Cureton, J. M. Porter, W. O. Jenkins, H. P. Emerson, and H. L. Parsons, I wish to extend my thanks. I owe to Dr. Cureton more than the usual obligation of a graduate student to his major professor for his suggestions and patience in his direction of this research. I also wish to express appreciation to Drs. Louise W. Cureton and E. O. Milton and to many others of the staff, faculty, and student body of the University of Tennessee for their help in the construction and printing of the scales used in the research.

The Secretary, Air University, United States Air Force; Mr. Prim of the International Business Machines office, Montgomery, Alabama; and Mr. James Reynolds all contributed materially during part of the data processing stage, and their assistance is gratefully acknowledged.

To all the teachers and students who participated, my sincere appreciation; I regret they must remain nameless.

I am, of course, indebted to many other researchers in area, but particularly to Mr. Roger Berkshire for his letters of advice and encouragement.

Eugenia Hill Cobb, my wife, has made this research possible by serving as typist, IBM punch card operator, and operator of a desk calculator. Her loyalty, energy, skill, and patience over the years can never be acknowledged fully by me.

Emmoran Benjamin Cobb

TABLE OF CONTENTS

CHAPTER PAGE

CHAPTER PAGE

LIST OF TABLES

LIST OF FIGURES

FIGURE PAGE

FIGURE PAGE

LIST OF APPENDIXES

APPENDIX PAGE

CHAPTER I

INTRODUCTION

General Introduction

This thesis reports the results obtained in applying various experimental methods of rating scale construction to the problem of teacher evaluation on the college level. An instrument produced through the use of such methods -- a scale to be used by college students in rating their teachers -- is presented and discussed. The major experimental method employed in construction of the rating scale, that of "forced-choice" (hereafter defined), was suggested by Professor Edward E. Cureton during a seminar in industrial psychology at the University of Tennessee in the fall of 1948.

Statement of the Problem

The problem, the construction of a rating scale for student evaluation of college teachers, had both "pure" and "applied" psychological research aspects. The "pure" facet of the problem was to develop, apply, and test the effectiveness of certain "forced-choice" principles of rating scale construction. The "applied" aspect was to construct an instrument useful to college teachers in the practice of their profession. Quotation marks are used for these terms to indicate the artificial nature of the dualism implied.

Importance of the Study

The need for this research was established in a number of ways before beginning work. A review of the literature established forced-choice as a promising method of scale construction, and correspondence with various departments of psychology indicated that the concept of applying forced-choice scales to the college teacher rating problem had merit and had not been done. A small-scale investigation of the actual use of scales in student rating of instructors revealed no use of forced-choice scales, some use of technically inadequate scales, and a large number of cases in which student rating of any kind was not attempted.

Need for the study was inferred through the assumption that construction of a better instrument -- one which could be widely used without regard to local conditions -- would contribute materially to effective study of the problem of instructor rating by students. The importance of research on teacher effectiveness as seen by students is supported by a large body of literature, and the importance of college teaching to our nation and civilization does not require delineation here.

Definition of Terms

To facilitate readability, the terms used in this thesis will be defined as they are used. "Forced-choice," however, is a major method and has been used above; this term is defined as follows: Forced-choice is a psychometric method employed in rating scale construction for the purpose of reducing the effect of rater bias. The method forces the

rater to select one, or more, statements about the person rated from a specific group of statements. The rater is instructed to select the most descriptive or best fitting statement. The group of statements has been organized on the basis of empirical information, previously obtained, so that all statements in the group appear to the rater as equally favorable (or unfavorable in one variation of the method). However, the group of statements has also been organized empirically so that one statement (more than one in some variations) is significantly better as a predictor of effectiveness than the others. The scale is then scored by counting the number of times the rater has selected "the nice thing to say" statement over against the number of times he has selected that statement which is equally "nice" but which also best discriminates between effective and less effective performance. Perhaps the method is best understood in terms of everyday experience by use of a crude analogy: that of "damning with faint praise," when, in conversation or letters of recommendation, the conventionally acceptable but unimportant traits are mentioned and the meaningful ones are glossed over or omitted.

Scope and Limitations of the Study

The research was directed to the problem of studying the effectiveness of methods and producing a useful scale for student rating of instructors in colleges and universities in the United States. It employed a large national sample: 70 schools, 124 classes, and 3,600 students in a

first sample; and 9 schools, 71 classes, and approximately 1,900 students in a cross-validation sample.

Early in the conceptualization stage of the research, various factors which might effect the results or be interesting in their own right were recognized. Such factors, therefore, as sex of teacher, whether the school was coeducational or male or female, age and academic rank of teacher, etc., were included as variables in the collection of data in the first sample. Such provision was made in the hope that this thesis might become the first phase of a program of research in college teacher evaluation and training. Because of the very large costs in time and money involved in this first phase, none of these variables have been studied. Part of the data have been punched on IBM (International Business Machine) cards used and are available to support future studies.

Although the scale as produced can be used effectively, various cautions stated throughout this thesis should be noted and observed in administering it on any given campus. Observations were limited to white schools, the sample is not known to be completely representative and unbiased, and reliability studies are yet to be made.

Methods of Procedure and Sources of Data

After the small scale investigation mentioned under Importance of the Study had been completed, it was decided that a study was feasible and would make a contribution to knowledge in this field of research. The literature was studied further and various instruments were designed to collect student beliefs about effective and ineffective college

teaching. These instruments were then tested, revised, and administered to various classes at the University of Tennessee. Administration of these instruments ceased when additional classes added few or no new concepts to the pool of ideas already collected. Through study of the data collected, a list of student likes and dislikes, opinions, and classroom experiences was prepared and used in the preparation of pilot scales.

Pilot scales, called the Qualification Check List, built on the above basis were administered at the University of Tennessee and the Georgia Institute of Technology, Atlanta, Georgia. After necessary revisions, these scales were printed and administered to a sample of colleges and universities throughout the United States. By punching the data on IBM cards and processing it, correlations were obtained between each of 394 items and two criterion scales, "effectiveness as a college teacher," and "liking as a person." Using these and related statistics, a validation scale was constructed and administered to a cross-validation sample under operational conditions. The forced-choice scale was then constructed on the basis of data in both samples. Various scoring formulas for the validation scale were developed and comparisons made, when the validation scale was used as if it were a forced-choice scale and when it was not.

Organization of the Study

Chapter II of the thesis presents the problem of student rating

of teachers in higher education and includes a discussion of the purposes, limitations, necessary assumptions, and advantages and disadvantages of formal types of student rating of instructors.

Chapter III discusses rating scale construction as such, with emphasis on forced-choice theory, method, and practice. The major objective of this chapter is to discuss the various technical steps taken in the construction of the instruments used in this research.

Chapters IV, V, and VI deal with the development of the Qualification Check List, a rating scale used to obtain information necessary to forced-choice methods of rating scale construction.

Chapter VII presents the reasons for the development of a validation scale, the methods used in constructing it, and steps taken in administering it.

Chapter VIII describes the results obtained from the administration of the validation scale and the results obtained from various methods of scoring it.

Chapter IX deals with the steps taken in construction of the forced-choice scale in the context of results obtained in cross validation and the requirements of further study of the problems involved.

Chapter X integrates the results obtained with proposals for further research. Various central problems of rating scale construction are used as a context for the discussion, and the results of the thesis are then summarized.

CHAPTER II

THE STATUS OF TEACHER RATING IN HIGHER EDUCATION

Introduction

An instrument or tool cannot be constructed or used effectively in a vaouum of purpose or without consideration of the social and psychological setting of its use. This chapter provides a brief analysis of the purposes and problems reported in the literature on student rating of college teachers; other kinds of teacher evaluation (for example, by administrator, peer, or self,) are not discussed except peripherally.

Estimate of Amount of Use of Student Ratings

In the small scale investigation mentioned in Chapter I, some thirty-seven letters were sent to Deans of Faculties. In this particular sample, about 30 per cent had used a scale or had attempted to do so, at the departmental level at least. Scales collected in this investigation were found, typically, to be brief and to contain either a narrow sample of quite specific statements or a number of "global" or highly generalized ones.

Gilliland (7), in a letter to member deans of the American Association of Collegiate Schools of Business, reported that thirty-eight of fifty-nine schools responding to a memorandum had no student rating of instructors. Thus the same rough order of magnitude applies -- about one-third had some form of student rating.

Dean L. S. Woodburne (31) visited forty-six colleges and universities and reported:

> One of the unfortunate conclusions to which the writer was forced was that the serious attempts to judge teaching effectiveness could be counted on the fingers of one hand . . . . Perhaps more discouraging than the scanty attempts at such evaluation was the widespread lack of concern with the problem.

Thus it appears that if Dean Woodburne's use of the term "serious" is taken seriously it is necessary to revise downward an estimate of 30 per cent use. It is to be noted that the quotation refers to evaluation of any kind, not student rating alone. It is possible to place this in context, if work done some fifteen years earlier can be assumed to be relatively stable. Anna Y. Reed (21) in 1932 surveyed 406 colleges and universities to obtain, among other data, administrator opinion on sources of information for evaluating teaching efficiency. When ranked among eleven items of teacher evaluation (by weighted value on a five-point scale from "greatest" to "no" value), "rating by students" had rank eight in the judgment of arts college administrators. "Rating by head of department" was most highly valued with rank one, followed by: two, "rating by dean;" three, "personal interviews and casual contacts;" four, "judgment of colleagues;" five, "rating by co-workers;" six, "student opinion;" and seven, "comprehensive examinations for seniors." "Surveys and observation by outside agencies" was ranked eleven, in last place. It would appear, then, that if little teacher evaluation is going on, and if value judgments of administrators, as reported by Reed, are effective determiners of college-wide activity, there can be very little effort expended in student rating of teachers on the college level.

This is not to say that efforts have not been made in the past or that outstanding exceptions to the above generalization do not exist. The University of Michigan, Brooklyn College, the University of Buffalo, Albright (Pennsylvania), Loyola (Illinois), University of Idaho, University of Illinois, University of Minnesota, and others have done important work in student rating. Nevertheless, the survey of the literature made for this thesis forces agreement with Medalie (17), who says, "Despite the success of faculty rating systems throughout the country, it has been the feeling of the National Student Association that the program to date is not enough."

Need for Evaluation of College Teaching

One possible explanation for the lack of wide-spread student rating of teachers is that there is no need for evaluation of teaching effectiveness. Algo D. Henderson (9), speaking as the analyst in the report of Group V (Evaluation of Teaching Effectiveness) at a conference in Chicago in 1950 sponsored by the American Council on Education and the United States Office of Education, in addition to describing the subject as "touchy," "controversial," and "important," stated:

> Generally speaking, faculty members of colleges and universities seem not to feel the urge to have evaluations of their teaching effectiveness made. While surveying the literature on the subject, I got the impression that there was a flurry of interest during the depression years. The literature seems not to be so extensive in subsequent years.
>
> . . . . . .
>
> College administrators, on the other hand, are caustic in their comments about the lack of interest in this important field. They are especially concerned that so little is being

done by way of experimentation and research.

It appears impossible to specify the amount of "good" and "bad" college teaching in quantitative terms and there is little need to do so in any event. The goal of improvement of college teaching does not necessitate description of a starting level; need for evaluation, furthermore, does not depend solely on an improvement basis. Maintenance of effectiveness is also a relevant and defensible goal. The necessary condition is that of establishing that evaluation methods do in fact lead to either maintenance of good teaching or improvement of teaching. A further quotation from Henderson (9) is pertinent here: "There is good evidence that most of the faults of poor teaching . . . can be improved upon when the individual has defined the nature of his inadequacies." The important question then, regardless of current opinion among either faculty or administration, is that of how to build and use effective evaluation instruments and methods. By definition, an evaluation method is effective to the degree that it facilitates desirable change. It is to be noted that stress should be laid here on the word method, in the sense that a useful evaluation instrument, in contrast, may be (and all too often is) inappropriately or wrongly applied. On the other hand, a defective instrument vitiates the evaluation method or process. Since a method must be applied in specific socio-psychological settings, this aspect of evaluation of teaching effectiveness becomes a matter of research and development on each individual college campus. This thesis work has produced an instrument, not a method. Furthermore, it is maintained that the instrument produced should be adapted, in use,

to such settings.

Methods and Purposes of Evaluation

It would appear that if the role of the college teacher has definite sub roles, such as teaching, research, administrative activities, etc., then a method effective in evaluating success in each should be developed. The problem of combining results from each method into an index of worth of the instructor to the institution is an administrative problem excluded from this thesis. It is not implied that the problem is unimportant; one source of fear and distrust of student ratings springs from the mistaken idea that such ratings can in some way substitute for an answer to this problem. Also, when an inadequate solution, a solution in conflict with the beliefs of the faculty, or a single rigid formula is applied, then the benefits of an effective method for one sub-role, especially that of teaching, are materially reduced.

Henderson (9) asserts that, "there is widespread lack of confidence in the usual administrative techniques of evaluation." He cites a 14 per cent "yes" reply to the question, "Do you believe that administrative officers have adequate information about teaching efficiency?" in evidence, as obtained in a 1944 study at the University of Washington. It is against this background that a central question as to the purpose of faculty evaluation must be raised. Should the purpose be to provide a personnel tool for use in placement, promotion, etc. or should the purpose be to assist the instructor to improve? The assumption underlying this question, of course, is that it is not feasible or desirable to develop and use a

single method to attain both objectives.

In considering this question in connection with rating as one method of evaluation, it is necessary to call attention to the fact that the question is not that of whether evaluation should be made; that is inevitable. The question is how to obtain all the relevant data in reliable fashion; and how to use the data effectively in achievement of objectives. Henderson (9) presents the situation in this way:

> There is no escaping the fact, however, that department heads, deans, and presidents must, and do, make evaluations. Should the judgment made by these officers be based on campus chatter, casual observations, and hasty impressions? Should it be based on the flow of complaints that cross every administrator's desk? Or should the judgment be the result of professional evaluation in the results of which the college and the instructor are mutually interested? I see no point in making two such evaluations. To be valid, they would have to cover the same ground and draw upon the same opinions; to be worth while they should both use the best available techniques. The end objective of each is to provide effective teaching.

The position taken in this thesis is that the state of affairs necessary to administrative participation could occur on some campuses. For this and other reasons to be discussed in later chapters, in administering a validation scale no attempt was made to restrict its use to that of educating teachers on the state of student opinion in their classes.

Causes For Lack of Use of Student Ratings

It has been established, then, that there is need for evaluation of the teaching process on the college level and that there has been relatively little activity reported in satisfaction of this need. The relationships between method and purpose and the administrative use of

ratings have been introduced. These two aspects, however, do not completely explain the disparity between need and activity. It is necessary to sketch in more background.

A useful lead in explaining the situation is found in Jacques Barzun's (1) distinction between education and teaching. Education, says Barzun, "is something unpredictable, . . . intangible, . . . something that comes from within a man's own doing;" whereas teaching is "something stable and clear and useful behind . . . education." Practical limits exist. "You know by instinct that it is impossible to 'teach' democracy or citizenship or a happy married life." Barzun admits that these educational objectives are connected to good teaching as by-products.

A discussion of the utility of this unfortunate point of view is not germane; the point is that many instructors may share it, and in sharing it believe that student rating (or any sort of rating) cannot evaluate teaching as education; the teacher is not responsible for it.

But it would seem that this distinction would serve to focus attention on teaching rather than education. How then can the meagre attention paid to student rating be explained? Barzun has a partial answer: "It is obvious the relation of teacher to pupil is an emotional one and most complex and unstable besides." Gilbert Highet (10) expresses this attitude more at length in these words:

> Teaching is an art, not a science. It seems to me very dangerous to apply the aims and methods of science to human beings as individuals, ... Teaching involves emotions, which cannot be systematically appraised and employed, and human values, which are quite outside the grasp of science. . . . You must realize that it cannot all be done by formulas . . . .

The destructive impact of this is no doubt rather obvious, but O'Hara (19) has expressed a possible and probably common product of this attitude so well that it cannot be dispensed with here. O'Hara (19) says:

> It is always a dangerous thing to put the label, Artist, on a man whose work you admire. In your enthusiasm for his work you may be doing him quite the opposite of a favor. After a man has been called an artist often enough (or too often), he begins to think he can do no wrong; that his entire work is Art and thus instinctively and automatically right, above criticism, and capable of understanding only by the elect.

Then, of course, there is that famous question of Rudyard Kipling "It's pretty, but is it Art?"

Another issue underlying the use of student ratings hinges on the function of the college and the role of the teacher. To quote Barzun (1) again, "If a man regularly exerts a positive influence on thirty-five out of a graduating class of three hundred and fifty, it makes no difference whether the remaining three hundred and fifteen loathe the sight of him . . . ." Riley et al (23), however, point out that ". . . the conservative view of the college as a place for the few has become an anachronism." Riley and his fellow authors stress the fact that this is a period of flux and change, of "new secular motives," of differing interests, range of abilities, and backgrounds, and of large numbers of students, and conclude that the concept of "the learned professor venerated solely for his knowledge is largely a phenomenon of the past." It would seem, however, that Barzun's attitude rather than Riley's is more widely shared by college teachers; if so, the teacher talking to "the students capable of learning the subject" would naturally find ratings by the entire class objectionable or at best mildly interesting.

A number of additional and interrelated "causes" for lack of use of student ratings are to be found in the literature. It appears impossible to generalize across all campuses and it is very difficult to partial out the rationalizations. The concepts included here are to be regarded as aspects worthy of some consideration, either in future research or in applying a student rating scale in a given department or campus.

Franz Schneider (27), for example, deals with academic freedom as a cause in these terms:

> Our present attempts at academic reform will fail because those sitting in our high academic councils lack the necessary information for effective action as well as the necessary organization for obtaining it.
>
> This is due to the secrecy which surrounds all classroom activities, the product of an obsolete -- and now very evil -- "academic sovereignty." Although in centuries past the promulgators of this "academic sovereignty" high-mindedly fought at great personal risk against the encroachments of an absolutist dynastic State or a dogmatic authoritarian Church, or both, "academic sovereignty" now serves but too often as a protecting cloak for shabby indifference and inexcusable waste.

However, Riley and colleagues (23), while recognizing that the professor is "situationally and psychologically sheltered" from the influences of student judgment because "the vagueness of definitions of goals or values of higher education preclude standards of judgment," and because "students are deemed incapable of good judgment," point to an interesting corollary in such a situation. They believe that this "closed system of protective devices built on isolation of the classroom . . ." surrounds men and women who, as teachers, approach teaching as a way of life, with a sense of dedication. The position taken in this thesis is that this sense of dedication or way of life is a powerful

force. A very small per cent of the faculty of most institutions would be unable to secure other employment more rewarding financially. The point under discussion, however, is the relationship of this dedication to student rating; namely, that "the way of life" would be intolerable if the teacher did not believe himself to be a good teacher. Even modifying this to include the teacher with the attitude, "I do research and present facts in class for those who can appreciate them," it remains true that the teacher must be satisfied that he is contributing. As Riley points out, it is easy then for this to lead to a rationalization that he is excellent! In such case, student ratings may constitute an unwelcome challenge to necessary but perhaps fragile beliefs.

A case history cited by Luella Cole (5) ties together the latent or sub-surface insecurity of teachers with student ratings. The seniors of a small college wished to honor their outstanding teachers by engraving their names on a tablet. Unable to agree, they consulted the president, who recommended rating scales. A joint student-faculty investigation was carried out, two names were engraved, and the affair forgotten by the students. However, says Cole, "The 'local kudos' within the faculty was more than anyone had anticipated." The two teachers were visited, teachers talked over their ratings with each other, and "there was a great flurry of experiments, projects, and revisions." The significant point, however, to the sense of dedication concept is contained in this statement of Cole:

> It is, the faculty members explain, not the sight of their names on the tablet that gives them pleasure but the knowledge that they are successes in their chosen work. To the sophisticate

> the whole thing may sound childish and silly . . . . In the college here described, however, the matter is taken not only seriously but almost reverently by both teachers and students.

Having treated, briefly indeed, some of the administrator-centered and teacher-centered sources of the problem of neglect of student ratings as a method of improving college teaching, it is now necessary to turn to a third source -- the student. Medalie (17), (Vice-President, United States National Student Association, 1950,) has this to say:

> . . . the formal interest of students in the educational process is itself an event of comparatively recent occurrence; however, with the new concept of the college as a community in which the welfare of the faculty, administration, and students is the common concern of every member of that community, the isolation of those who are most immediately affected by education -- the students -- is gradually becoming a thing of the past.

The fact that the college-as-a-community concept is hardly new need not detract from realization that its existence largely as an ideal rather than a working pattern detracts from student awareness and interest in ratings. In any event, student interest can by no means be assumed. For example, Medalie says, "One of the greatest weaknesses . . . has been the failure to promote direct student action and interest in the further analysis of the curriculum and the educational process." Medalie also quotes the Harvard student council report as saying, "It is for us (students) to change the general attitude from one of indifference to one of active participation . . . ."

It should be noted that the reference here is to <u>formal</u> participation, an active-working-at-it-systematically sort of thing, rather than student awareness and interest in teaching as recipients. Experience with student ratings clearly indicates that relatively little effort is

needed to activate and channelize successfully student interest in teaching.

It is against this background that specific arguments in opposition to student rating of teachers should be evaluated. They are included here as products of the attitudes described, and as factors to be considered in developing methods of evaluation. Bryan (4) has listed these arguments as follows:

> 1. Students are not competent to judge the merits of the teaching process or the results.
>
> (a) "It is difficult for the learner to differentiate between indoctrination and good teaching. The ideal teacher is that teacher who so trains his students that they become increasingly independent of his instruction and guidance. The result of having 'too much teacher' may be either over-coddling or over-indoctrination. From the very nature of the case, the student who is being over-coddled or over-indoctrinated is least likely to be aware of what is taking place." It is thus that Lehman (15) states what he considers "one of the most serious objections to student attempts to rate faculty members."
>
> (b) "Teacher-activity is of value only to the extent that it brings about well-directed pupil activity. The critical factor in successful teaching is thus not what the instructor does; it is what he succeeds in getting the pupils to do. This principle that education is an outcome of self-activity holds for all levels of instruction from the kindergarten to the graduate school of the university. If one may judge from the attitude displayed by many undergraduate students toward study, it seems likely that the principle of learning through self-activity is not as generally understood among students as might be desired. Obviously a student body will need to be carefully instructed in this matter before faculty rating is attempted." (15)
>
> (c) "The persuasion that teaching is best which pleases the majority of students is surely a most glowing example of 'democratic' fallacy. If the teacher's promotion is to depend upon student ratings, student prizes should be awarded to those who graduate precisely in the middle of the class, and the artist should be given an award only if his painting is judged to be best by the majority of those who visit the exhibition. The above statement summarizes the position of the present writer (Lehman (15)), who firmly believes that too much emphasis placed upon student ratings of their teachers may defeat the very end pursued, namely, the improvement of instructional efficiency."

(d) The judgment of students is immature. Young people are given to snap judgments. Their first estimates may not correspond with those given a few months or a few years later.

2. The validity and reliability of student ratings may be further affected by one or more of the following influences:

(a) Low grades.

(b) Fondness or dislike for the teacher.

(c) Amount of work required by the teacher.

(d) Interest in the subject. A student interested in a subject may rate the teacher of that subject high, even though such a rating is not warranted. Conversely, a teacher may be rated unjustly low because the student lacks interest in the subject.

(e) Difficulty of the subject.

(f) Reputation of the teacher. Beginning students who have not become well acquainted with a teacher may rate him high or low on the basis of what upperclassmen have said.

(g) Dislike for, or boredom with, too many ratings. There may be, in such cases, a tendency to rate uniformly high.

(h) General attitude toward the school. If a student has an unfavorable attitude toward the school as a whole, it may be reflected in the rating of an individual teacher.

(i) Wrong attitude. Students may treat the assignment of rating of teachers as a joke.

3. Permitting students to rate teachers may have disruptive effects on the morale of the faculty.

(a) The hostility of teachers to being rated by students may interfere with teaching efficiency.

(b) In case a teacher's rating turns out to be rather low, he may become seriously discouraged.

(c) Ratings may make teachers self-conscious.

(d) Young teachers may lose respect for elder teachers.

(e) Teachers may cater to student opinion. This may lead to directing efforts to obtain popularity with students through outside-of-class activities at the expense of the quality of teaching within the classroom.

4. Permitting students to rate teachers may have disruptive effects on the morale of the students, which may take the form of one or all of the following:

(a) Weakened respect for teachers and superiors generally.

(b) A feeling that students are the judges of the value of teachers, curriculum, and content of courses.

(c) An expectation that teachers should change their ways and plans as per ratings.

5. Having students rate teachers is costly in both time and money.

Results of Student Rating

Perhaps the best way of evaluating these and similar arguments is to consider the results obtained when student rating is used. The literature from 1900 to 1952 has been surveyed by Morsh and Wilder (18). These authors summarize the literature on student ratings by saying that use of student ratings is growing, that they have fair consistency, and that when compared with other measures of effectiveness diverse results are obtained depending on criterion used. They also report that considerable halo is usually found, and that the effect of grades depends on the instructional situation. Size of class, sex, age, maturity, and intelligence seem to have little bearing. These authors also conclude that:

> Research has been too sporadic and results too inconclusive to allow generalizations to be made concerning the influence on student ratings of other factors such as age and sex of teacher, length of students' acquaintance with the teacher, length of time teacher has taught in the school or taught a student, pleasurable personal relationships between student and teacher, and whether or not subject taught by rated teacher is students' favorite subject. There is considerable expressed opinion but little research evident that student ratings will contribute to instructor improvement or could be used to improve supervisory ratings.

The President's Commission on Higher Education (20), in 1947, reported that:

> The case for student evaluation of instruction as a means for improving teaching excellence is a strong one. Most theoretical objections to it vanish when practical try-out is made; faculty acceptance is typically favorable.

Summary

It would appear that this last quotation is not supported by results reported in the literature as summarized by Morsh and Wilder (18). These quotations were selected deliberately as characteristic of two positions held regarding student rating of teachers -- one that of strong belief, the other a typical product of the scientific investigator. A third position, that of rather violent objection or an avoidance reaction, has been shown to be the majority point of view. This chapter has analyzed briefly the etiology of this last attitude.

There is a need for research in the social psychology of student rating of teachers. It would seem that generalizations could be derived from rigorously conducted research at any given institution which could be applied successfully on other campuses. To make such generalizations valid, however, the instrument used must be designed, not for local use alone, but for application on a large number of campuses. In view of the large amount of suspicion of student rating, this instrument should also be one which can be demonstrated to have minimum biasibility, along with appreciable objectivity. and utility. The research reported here was based on these considerations. The next chapter will discuss the psychometric considerations involved.

CHAPTER III

FORCED-CHOICE THEORY, METHOD, AND PRACTICE

Introduction

Various aspects of evaluation methodology as presented in the last chapter will be considered in the context of this chapter, which is one of analysis of the problems of development of a rating scale instrument. The major purpose of this chapter is to review the psychometric considerations involved in constructing forced-choice versus conventional scales, and to analyze rationale and methodology, so as to explain the procedures adopted in the development of the scale reported in this thesis.

Conventional Versus Forced-Choice Related to Purpose of Rating

For the purpose of this discussion, scales other than forced-choice, such as graphic, linear, descriptive, etc. are lumped together under the rubric "conventional." "Forced-choice," defined earlier, will be analyzed in detail later in this chapter and differences between the two types of scales will be elaborated.

The relationship between the purpose of rating and the utility of forced-choice versus conventional scales is expressed by Highland and Berkshire (11) in these terms:

> It is perhaps a legitimate criticism of the forced-choice method that the completed form is of little use to the supervisor in counseling his workers as to their strong and weak points. This may be an inevitable characteristic of any rating

> method that has as its primary purpose the provision of accurate information regarding the relative abilities of men.

These authors quote Rundquist and Bittner (24) as saying:

> Ratings which are to serve as a basis for administrative action must yield a valid measure of an individual's performance relative to that of other individuals. The rating system must be specially designed for this purpose; in such a system the rating form or scale takes on particular significance . . . . However, such procedures will probably be found to be of little use in assigning work, in raising morale, or in helping people to improve.

However, the position has been taken in this thesis that in considering an evaluation method an either-or choice between administrative use and teacher use is unnecessary. The construction of two separate rating scales for student use is quite clearly a likely source of confusion on the part of all concerned. Aside from many other theoretical considerations, the cost would be excessive.

Two solutions to this problem appear. One corrective measure is mentioned by Highland and Berkshire (11) as that of appending a sheet to the forced-choice scale on which the rater may "indicate in a systematic fashion the individual's strong and weak points." They point out that "a copy of the information recorded on this extra sheet may be retained and used by the rater in the counseling and training of his workers." Obviously, supervisor use of the scale is being considered here; no suggestions are offered on how to handle the problem of student rating of teachers.

Another solution was used in this research. Briefly, the method used was to construct the scale so that it gives the appearance of being nothing more than a graphic rating scale, yet it can be scored as if it

were a forced-choice scale. The graphic scale is used to "indicate strong and weak points;" a scoring key suppressing certain items, or some other key, may be used to obtain an evaluation index number.

Conventional Versus Forced-Choice Related to Research

In the summary of Chapter II it was said that research on the effectiveness of college teaching through the use of rating scales would require that scales be designed for use on a number of campuses; otherwise, validity generalizations could not occur. The conventional rating scale, in spite of its defects, does lend itself to such use. What is the situation with regard to forced-choice scales?

Seeley (28), writing in 1948, had this to say concerning forced-choice method:

> The technique is readily adopted, when properly handled, for use in any rating situation; the rating scales themselves, however, cannot be adapted -- or adopted -- for outside use. The reason for this lies in the very nature of the method (which also accounts for its success): it yields rating scales too specific for use in situations other than the one for which they were built.

However, the last phrase in this quotation, "other than the one for which they were built," provides a cue; would it not be possible to define the "one situation" as a common one across many campuses? In other words, the researcher could obtain a large number of specific items and eliminate those which do not apply across the board to all classes in the sample. Even though no feasible method was discovered for evaluating the success of this method against the method of construction of a series of locally-bound forced-choice scales, this method was adopted. The

criterion for inclusion of an item in the scale was, however, one cross-validation sample; it was not a criterion based on computation of item-correlations for each item by class and selection of only those items with consistently high correlations in all classes.

Conventional Versus Forced-Choice Effectiveness

Before turning to an analysis of differences between conventional and forced-choice methodology, it is perhaps wise to provide an orientation for the comparison. A statement by Highland and Berkshire (11) is so useful in this connection that it is quoted at length. These authors say:

> A word of caution seems necessary here. While the study reported in the following section of this bulletin has as one of its aims a comparison between conventional and forced-choice rating methods, it is difficult to make a fair comparison from the available research literature. The materials on the deficiencies of conventional rating methods presented above grew out of some 30 years of experience with them. Experience with forced-choice rating methods is very limited. The few published research reports on forced-choice rating tend to emphasize those aspects of it that correct the deficiencies of older methods. A review of these reports thus yields an unavoidable emphasis on the superiorities of the forced-choice method. It could well be that, as more is learned about the results of forced-choice ratings under operational rather than experimental conditions, some of these apparent superiorities will vanish. It is also quite conceivable that, as experience with forced-choice ratings accumulates, psychologists may find that the method has its own unique deficiencies.

Travers (30) has taken this position:

> An examination of the validation studies of forced-choice assessment methods reveals that the evidence does not support some of the claims made for the validity of these procedures. The results show that in studies in which little criticism can be made of the procedure, there seems little to choose between forced-choice and traditional rating.

Brogden (3), writing three years later than Travers, says that, "Usable validity has been well established in a number of questionnaires based on forced-choice in the prediction of success of military leaders," whereas "consistently lower validities are obtained with questionnaire items in yes-no form with conventional keying." Brogden adds that, "forced-choice appears to offer the most promising lead."

The necessity of both careful analysis of theory and research to test the theory is therefore apparent. The following sections of this chapter make such an analysis and relate it to the work done in this research.

Forced-Choice Practice

Perhaps the most useful method of analysis is one which starts with practice (as exposition of method) so as to provide a framework for a discussion of theory. What has been done will be presented, then why it was done in this way.

Also, the framework will be organized around what will be called here the "standard" method, as developed by the Personnel Research Section, Department of the Army, and reported by Sisson (29). "Standard" is used instead of, say, "original" to indicate that there have been a number of applications of this method by consulting firms in military and industrial settings. Highland and Berkshire's (11) "note of caution" as quoted earlier is mentioned to remind the reader not to equate "standard" with "accepted;" also, a number of quite radical variations have been used in building forced-choice scales. These will be discussed in this chapter.

The "standard" method, then, consists of a series of essential steps. These steps are numbered to facilitate reference in the discussion to follow; for example, "standard step 2."

Following Sisson (29), the "standard steps" are:

1. Descriptions are collected of individuals who are at each extreme of the scale to be measured. If the scale is that of effectiveness as a supervisor, descriptions are obtained of the most effective and the least effective supervisors. This procedure partly defines the scale that is to be measured by the final instrument.
2. The descriptions collected are then dissected into a list of small elements. Each one of these elements describes, in essence, a rather specific item of behavior. The complete list should cover all the important aspects of the job, and the number of items covering each aspect should be related in some rational way to the importance of that aspect.
3. Two index values are determined for each item listed. One, the discrimination value, indicates the degree to which the item measures the particular characteristic that is being measured. The other value indicates the extent to which individuals tend to rate others high or low, generally, on the particular characteristic. This latter is sometimes referred to as the preference value of the item. Both values are determined experimentally.
4. The characteristics are then arranged in pairs such that the two members of each pair are equal in preference value but differ in the extent to which they discriminate. Ideally, one of the characteristics in each pair should have a discrimination value of zero and one should have a high discrimination value. Also, it should be impossible for those who are to use the scale to determine by inspection which one of the two characteristics is the discriminating one.
5. The pairs of characteristics may then be grouped in fours, with each group of four including two desirable and two undesirable characteristics. The main purpose of grouping the characteristics in groups of four seems to be that persons filling out the scale may have a better attitude towards it if the grouping in tetrads is adopted; because, of the two negative statements, raters are only required to check the one which is least like the rater. This seems to be the only advantage of the tetrad over the couplet form. A pentad form has also been suggested.
6. Directions are prepared in which the individual is instructed that he is to examine each group of four

characteristics and to select the one that is most characteristic and the one that is least characteristic of the person who is being rated. The person filling in the form is required to make these choices.

7. The selection of the items is then validated against an external criterion on a sample that was not used in the original procedure for selecting and pairing the items.

Forced-Choice Rationale

Treating first the "standard" practice as quoted above, the rationale of using forced-choice format -- that is, requiring the rater to choose the most characteristic item rather than reporting the degree of possession of the quality as in a graphic scale -- is regarded by Richardson (22) as basic to forced-choice method. In forced-choice, he says, the rater is primarily reporting the job performance of the ratee, not evaluating it. Evaluation in forced-choice is accomplished by empirically established scoring methods. The point is made that the conventional scale is defective in requiring both at the same time because such attempts to evaluate bias seriously the reporting effort.

A second aspect of rationale, according to Richardson, is that the forced-choice format forces a critical judgment, at least one over and beyond that produced by the conventional scale, in that the rater is forced to think back to specific events and cannot depend on his general opinion or idea of the ratee to choose the most characteristic item.

An analysis of this position will take the simplest case, and the case used in the research, that of a couplet of two positive or

favorable statements; for example:

1. Has good posture and bearing.

2. Uses sufficient supporting details in lesson.

Following Travers (30) the question is asked, "What does the rater do in completing a forced-choice rating?" Travers maintains that, far from reporting job performance, the primary task of the rater is to decide upon some standard or frame of reference (for example, the average posture and bearing of teachers he has known,) and then evaluate where the teacher being rated stands with respect to this standard. Only after the rater has placed the ratee on a continuum for both items is he is a position to judge which is most characteristic. That is, the ratee might be above the average teacher in frequency or successful use of supporting details, and around average with respect to posture and bearing, in which case item 2 would be chosen as most characteristic.

One could perhaps argue, in view of the time used in completing scales, and on the grounds that much of human behavior is non-logically rather than logically oriented, that choice of the most characteristic item just does not occur in this fashion. Then, how does the choice take place? One possible response to a forced-choice couplet is that of a global or "I remember him best as" action. This sort of analysis could quickly become confounded with a discussion of "images," of "stereotypes," and various other psychological constructs. However, the literature appears to lack any research specifically pointed to these questions; and since the method does appear to work, it appears more profitable to explore answers to the question from a psychometric point of view

rather than to speculate on what goes on in the rater's mind. It is recognized, of course, that research on this point might have value in improving the method further.

Before turning to the question of psychometric answers as to why forced-choice works, it should be noted that the global concept above is not consistent with Richardson's (22) assertion that, in forced-choice, the rater is forced to think back to specific events. Thus Travers's (30) comments cannot be dismissed or lightly regarded and the necessity of research on this point becomes apparent.

Brogden (3) has contributed an important concept on the psychometric reason for forced-choice success and has proposed a way of improving the method. Retaining the simplest case, the forced-choice pair or couplet, Brogden holds that forced-choice works because that part of the variance in the score due to distortion (i.e. unreliability due to bias) is reduced by suppressor action when scales are converted to forced-choice. The forced-choice pair is regarded as a <u>difference score</u> in which, according to Brogden, the less valid item (the one with lower correlation with the criterion) has "unit negative weight." The more valid item alone then contributes to the score and the invalid item acts as a suppressor. Thus distortion variance is reduced.

It should be noted that regarding the pair as a difference score supports Travers's analysis of the psychological situation facing the rater. It also introduces the question, if this is true, whether or not the forced-choice format is necessary to increase validity of a rating scale. Travers maintains that it is not necessary to ask

the rater to choose the most descriptive or characteristic item of a pair. The scale can be presented in graphic form, the score of each item obtained, and a difference score can be thus derived. Brogden (3) also makes this same point, saying, "It . . . should be remarked that the rationale and procedures could be modified and used without forced-choice format."

The problem, then, becomes one of examining "standard" forced-choice methodology to see if it is adequate to a suppressor variable approach. Before doing this, however, another aspect of the question of forced-choice format deserves consideration. As Rundquist (25) pointed out back in 1946:

> However, another method used by the Personnel Research Section yields equally good validity. This method involves the use of separate ungrouped items selected on the basis of the discriminative indices described earlier. In the presentation of these the rater simply indicates on a five-point scale the degree to which each descriptive phrase applies to the person rated.

What, then, is the advantage of using forced-choice format? Rundquist answers that, "the logic of its construction offers more promise of obtaining results less subject to rater control than any other known method." Travers disputes this, saying that:

> If a rater decides he wants to give a person a higher rating for job proficiency than is deserved, all he has to do is to think of the person who was generally considered to be most proficient in that job and to fill out the forced-choice form for that person. The technique would almost certainly ensure a high score. In a similar way a low score could be assigned at the will of the rater. The method which a rater has to use to control the rating is a little more complex with the forced-choice technique than with the conventional rating technique. It is still an open question as to what fraction of raters will see

this loophole in the technique, but loopholes of this kind are likely to become common knowledge.

It should be noted that this discussion has been concerned with conscious control. It may be, as Richardson (22) maintains, that tendencies to over-rate and to show bias are deep seated and unconscious, and therefore they must be counteracted. Some very interesting questions present themselves at this point, however, if student rating of college teachers is maintained as the context for the problem. These questions arise from the nature of the relationship between the purpose of rating and its validity. The position taken in this thesis has been that the purpose is neither administrator nor teacher centered; that the purpose is to present as many meaningful and reliable statements as possible about student response to the teacher's behavior and to the classroom life and atmosphere. The bias that Richardson (22) mentions becomes completely unmeasurable and a meaningless abstraction when a class concensus on a given item can be clearly and consistently obtained. "The validity" of the scale here <u>is</u> its reliability. To the extent that "bias" is consistent, it can no longer be bias. However, any instrument has as many predictive values as criteria predicted, and this is not to gainsay that correlation of scale results with other criteria, such as peer judgment, gain in subject matter knowledge, etc. could be quite low. This does not appear to be a problem which can be resolved apart from considerations of relevance. As stated earlier, the teacher has many roles, and an attempt to evaluate these, in combination, with a single index number, appears futile. As far as conscious prejudice is concerned, it would appear that one or two individuals

in a class of usual size could not distort results markedly; if there are "too many" of these, an inspection of the distributions should be rewarding -- otherwise such bias is part of that which should be measured.

A complementary question, what are the disadvantages of forced-choice format, has been answered by Highland and Berkshire (11) in terms of rater resistance. Considerable resistance was found in Army use. The disadvantage of forced-choice in improving instruction was discussed in section one of this chapter.

One further aspect of the problem, bias, and particularly leniency, should be associated with the way ratings are to be used. In student rating of teachers, it is almost mandatory that ratings be unsigned. Thus it appears that forced-choice format would not be particularly useful in reducing student tensions or in obtaining more reliable evaluations.

Regardless of the position taken up to this point, a further question arises, can the basic rationale (defining it as suppressor variable in nature) be applied in practice without forcing a judgment? We have seen that Brogden (3) states that this is possible.

In summary, then, it appears that forced-choice format is both unnecessary and in some cases undesirable. For these and other reasons to be discussed in the next section, the validation scale was constructed and administered in graphic format.

Suppressor-Variable Rationale

Brogden (3), having defined the essential rationale behind forced-choice as that of suppressor action, goes on to point out that in techniques in current use forced-choice scales "are not explicitly oriented to get suppressor action." The reason, according to Brogden, is that current usage tries to eliminate distortion by use of two items which have equal <u>average</u> (italics Brogden's) ratings on social desirability. Individual differences in tendencies to distort are lost in computing these averages and they "bear no explicit relationship to the reduction of invalid variance."

In the remainder of his article, Brogden goes on to develop a rationale and procedures for producing pairs of items free of distortion, so that correlation with a "distortion score" is zero. This procedure is significant for the program of research envisioned as a result of the research described here and, as such, is treated in Chapter X. The pertinent point at the moment is the limitation recognized by Brogden and expressed in these words: "but it must be assumed that the correlation properties of the alternatives forming the pair are unaltered when these alternatives are presented as pairs." The decision to administer the validation scale in graphic rather than forced-choice format in this research was taken several years before publication by Brogden, but for this same reason. That is, accepting bias as a problem, if the item properties are known, then a graphic scale can be keyed and scored to eliminate bias without making such an assumption. It is at this point in his article that Brogden points out that forced-choice format is not

neoessary.

Methods Used in Construction of the Scale - Preference Index

Various departures were made from the "standard method" as outlined by Sisson. The rationale behind these departures is presented in this section. A detailed account of the procedures used is to be found in Chapters IV, V, VI, VII, VIII and IX.

In standard steps one and two, in this case collection of information on good and poor college teaching and isolation of behavior elements, no important change in rationale was made. As reported in Chapter IV, several different methods of data collection were employed.

In standard step three, the discrimination index was obtained by computing product-moment correlation coefficients for each item in a Qualification Check List and a criterion scale, rather than obtaining indices with only "good" and "poor" groups of teachers. This, however, is not a shift in rationale but is, rather, a refinement in technique. The preference index, on the other hand, required a new approach. Rationale considerations here merit extended discussion; to avoid excessive length, and because Berkshire (2) and especially Lanman and Remmers (14) have already done so, an historical approach will be sacrificed. Instead, the concepts themselves will be discussed.

It will be recalled that the purpose of pairing items of approximately equal acceptance value socially, that is, of equal attractiveness, is to prevent the rater from consistently selecting the more discriminating item of each pair on the basis that it is also

more attractive. The total score in this case would be subject to rater bias, especially leniency. In the discussion here the assumption that if items are equally acceptable the scale is more resistant to bias is retained, in spite of Travers's (30) criticisms and Brogden's (3) statements. Although the validation scale was administered in graphic form, it was to be scored as if it were a forced-choice scale. This meant that the scale needed to be constructed according to forced-choice rationale, so that the only difference between the validation scale and a forced-choice scale would be in the directions given to the rater. At the time of validation scale construction, Brogden's article regarding suppressor variable rationale had not been published, so that the research was not oriented around this concept. However, it was recognized that no direct cross validation of items (as contrasted with couplet validities) could be obtained if the validation scale were administered in forced-choice form. In any event, the decision made was a fortunate one, in the sense that future research based on suppressor variable technique will not be restrained by the assumption that the correlations of the items in the couplets remain the same when these same items are presented as couplets. As pointed out in Chapter X, in future research it will be possible to apply Brogden's technique to the validation scale statistics already gathered. Also, it will be possible to test the assumption re item correlations by administration of both the validation and forced-choice scales to the same classes, half of the class receiving one scale, half of the class receiving the other.

Turning, then, to the problem of obtaining an index of social value

or attractiveness, the work of Berkshire (2) was found to be especially valuable on this point. In investigating this problem, Berkshire first calls attention to two necessary assumptions: first, that the items or statements "have a characteristic of attractiveness that is substantially independent of the ability of a statement to discriminate between good and poor workers;" and second, "that the particular index chosen to represent this characteristic of attractiveness is actually a valid measure thereof."

Berkshire conducted an experiment to test the first assumption and to determine which of several kinds of indices in use was best. The method used was to hold constant the discrimination index for a block of statements and to vary the attractiveness index. Raters were asked to give as high a score as possible. The hypothesis was that, since the discrimination index was held constant, any between-rater consistency of choice would reveal any independent statement-attractiveness characteristic. Correlation size between frequency of choice of statements and type of attractiveness index was used to determine which type was most valid.

Berkshire (2) concluded that a characteristic of attractiveness "at least partially independent of the ability of the statement to discriminate" does exist.

Berkshire tested five kinds of indices.

The first of these, the "preference index," has been described by Sisson (29) as "the extent to which people in general tend to use it in describing other people," as "general favorableness" and as "the

tendency of raters to mark people high or low." In operation, Seeley (28) used the mean score on the item as the preference index. Berkshire (2) concludes: "The preference index does not reflect any variance in statement attractiveness -- it is not only useless, it is pointless."

At the time the validation scale used in this thesis research was built, Berkshire's results were not available. However, it was recognized that the preference index was not truly a favorableness index and could not be relied upon as the sole basis of constructing couplets. Accordingly, it was retained under the name of the "characteristicness index" and used in conjunction with another index, the "index of biasiability," as discussed later.

The second index, developed by Highland and Berkshire (11), and called the "favorableness index," is derived by asking judges to rate each item or statement on a five-point scale which ranges from "very unfavorable" to "highly favorable." The mean of the ratings for a given item is used as the favorableness index. Berkshire concludes that this index "reflects so little independent variance in statement attractiveness as to be of little value."

A third index, called the "upper group applicability index," was also found by Berkshire to be invalid and "worthless in the construction of forced-choice blocks." This index amounts to using the mean applicability (i.e. preference) of the upper group. It is mentioned here because it was developed by Harding and Long (8) in an attempt to get away from the additional step in data collection required by the favorableness index, an objection shared by the writer but met by

developing the biasiability index, as described later on.

The other two indices were the "face validity index" and the "job importance index" as developed, according to Berkshire (2), by George Burgess but unreported by him. The face validity index is obtained by calculating means from administration of a scale in the form:

This statement could be made

About all instructors, including the poorest.

Of all but the poorest instructors.

Only of average or better instructors.

Only of better than average instructors.

Only of very good instructors.

The job importance index is obtained in the same way by using this sort of scale:

For the job of Air Force instructor I consider the behavior characteristics described by this statement to be

Of no importance.

Of minor importance.

Somewhat important.

Quite important.

Of greatest importance.

The face validity index, according to Berkshire (2), has little independent variance, whereas the job importance index "is the standout," and "is the only one of those evaluated which has the qualities necessary . . . it reflects a considerable amount of variance in attractiveness that

is independent of variance in discrimination."

In this research, a still different approach was taken. The basic notion arose from the student comment, "He's a heck of a nice guy, but he just can't get it across." Clearly, two criterion scales were indicated: one which would allow the student to express how he felt about the teacher as a person, and one to permit him to evaluate teaching effectiveness. One desirable feature of two such criterion scales was the fact that, in the collection of data, effectiveness as a teacher scale can be purified of some bias by putting the liking as a person judgment first. In this way the position is clarified for the rater; he understands that two separate and distinct decisions are required; and by discharging an emotional loading first, one either favorable or unfavorable, he is in a better psychological situation to rate teaching effectiveness. It is also possible, of course, that such clarification gives the student the opportunity to deliberately bias his effectiveness rating; or, that there is an unconscious but strong carry-over from one criterion rating to the other. Since these two criterion scales were administered under conditions of rating of an unnamed teacher in the Qualification Check List, and of rating the teacher in the Validation Scale, a check from this standpoint was possible.

Once a liking-as-a-person criterion scale had been decided upon, it was seen that the relationship between each item and this criterion could serve as an index of biasability. It seems clear intuitively that either conscious or unconscious bias results from the person-to-

person relationship, and that the item favorableness or the social acceptability of the statement is dependent upon this relationship. The correlation between the criterion scales in a pilot run of the Qualification Check List indicated that about 50 per cent of the variance in one scale could be accounted for by the other scale. Therefore, item correlations with these criteria scales could be expected to have some independence and the biasability index was feasible. The theoretical advantage to this index is that it is not contaminated by stereotypes and is direct rather than indirect. The student is not asked what he could or would do about a statement, and, therefore, the popular notion cannot intervene in his response to it. No assumption is required that he will react, in using a scale, the way he says he ought to react.

Methods Used - Content and Format

Returning to "standard step three," the determination of the indices, it is obvious that these statistics must be obtained empirically. The vehicle by which these data were obtained is called the Qualification Check List in this thesis and is described fully in Chapters IV, V, and VI. Certain general aspects, however, may be considered here.

In the previous section reference was made to variations in technique in connection with the kind of sample used. In "standard" practice, a group of "good" and "bad" individuals are used to obtain item statistics. The necessary assumption in such case is that the whole group is adequately represented by taking half the sum of the

mean score of "good" and "bad" groups -- that the two ends of a normal distribution have been obtained. This assumption was considered unsafe in the work reported here. The specification of a "best" or "poorest" group, moreover, increases the possibility of halo operating to reduce item validities.

Another decision made early in the work was to avoid unfavorable or negative statements. Rundquist (25) had pointed out that raters tended to object to even mildly unfavorable ones. In almost all cases it was found possible to restate negative items in logical opposite fashion. It is to be noted that Highland and Berkshire (11) report that of six different kinds of forced-choice blocks tested the one in which all statements had a high favorableness index was the one which gave the best overall results. One of their two general conclusions is that "the inclusion of both favorable and unfavorable statements in the same block appears to be an inferior method of constructing forced-choice forms." The couplet form was employed rather than the triad or tetrad form (three or four statements in a block), primarily in order to use two control indexes, those of characteristicness and biasability, and secondarily because this was a variation not tested by Highland and Berkshire.

Summary

It will be apparent that forced-choice theory, method, and practice have been evolving rapidly since 1946 and that many questions remain unsolved. The major question appears to be whether standard

forced-choice format is necessary or desirable to achieve the ends for which forced-choice was designed, especially in its use in student rating of college teachers. A general background has been presented as context for the procedures and results described in later chapters.

CHAPTER IV

THE PREPARATION OF THE QUALIFICATION CHECK LIST

Introduction

The preparation of the Qualification Check List was considered to be one of the most important steps in the construction of the scale, and various methods were used to insure adequate coverage in producing a pool of phrases descriptive of college teachers. Some of these attempts failed, but are included as information for those who may attempt to use similar methods.

Analysis of the Literature

Texts and articles in the fields of education and psychology were surveyed. Two factors soon caused curtailment of this method. The first was the concept advanced by Rundquist (25) that forced-choice scales should consist of phrases derived from the notions and language of the raters. The rationale behind this is one not only of ease of reading or acceptability by the student, but is also a matter of ratability. If concepts advanced in the scale are new to the student, then it is quite possible that he may fail to understand a complex idea presented in a short phrase. Furthermore, if it is a new concept, it is probable that he has no background of observation of the teacher from which to evaluate. This applies, of course, to each rater; so that carried to its logical limits, all students should be familiar with all the phrases

used. In the Qualification Check List (called QCL hereafter), as finally produced, this probably does not hold true, but provision was made for elimination of non-meaningful phrases in the final scale by use of Column 6 in the QCL and later analysis of the frequency of checking it.

Concepts in the literature were used, then, primarily as ways of recognizing implied student comments or judgments when student essays or other records were read. It should be noted that some editing of student contributions was necessary in order to condense ideas into short phrases and to put negative phrases into positive or favorable form. An effort was made to avoid distorting or altering student ideas, but there was some departure from student notions and language and perhaps some sacrifice of meaning.

Faculty Interviews

Before the considerations above were fully understood, a work sheet for interviewing faculty members was mimeographed, and a few tentative interviews were made, with the idea that valuable information might be obtained. This method was inefficient. Given sufficient resources, perhaps the interview method with both student and faculty is desirable. This is particularly true if the "critical incident" technique, as presented later in this chapter, is to be used.

Faculty Incomplete Sentences Questionnaire

As a substitute for the interview method, a questionnaire was devised and sent to the entire teaching staff at the University of Tennessee by campus mail. The return was small, but some usable concepts were collected. This lends some force to the desirability of using the interview method.

It was assumed that many teachers had given years of thought and study to the problem of teaching and that there was no necessary conflict between their conclusions as a group and students' conclusions as a group. Because of the failures cited, however, a clear-cut listing of such teacher conclusions was not available to compare with student concepts. The faculty aspect in building a QCL was therefore abandoned.

Collection of Descriptive Phrases from Students

Essays were collected from the writer's classes, and the phrases obtained were written on 5x8 cards and filed under a preliminary organization scheme as shown in Appendix A. These rough categories were used to check phrases for duplication as new student records were read and to organize the final forms of the QCL.

An attempt was made to use the idea of change in student opinion as a source of descriptive phrases. Students were asked to furnish their first impressions of teachers new to them by completing a blank at the end of the first week of class. About two weeks later another blank was furnished, and students were asked to confirm or change their ratings

and to describe the incidents responsible in either case. Forms were placed in about every third "distribution box" at the University of Tennessee. Returns were disappointingly small; but the few dozen forms received were returned two weeks later via distribution box, and the additional information gathered was filed in the item pool of 5x8 cards. The form was long, cumbersome, and the method of collection was inadequate.

A large number of items had been collected at this point, but the writer was not satisfied that student opinion had been thoroughly studied. Some years earlier the writer had worked on an airline pilot study for the American Institute of Research and had become familiar with an interviewing method called "the critical incident technique," as reported by Flanagan (6). The essence of the approach is the notion that specific critical events are apt to be good sources of information concerning an activity; that "critical incidents," when explored with a subject, produce observations rather than vague generalizations. The critical incident may be defined by the observer, as in "the last accident or near miss you were in as a pilot;" or the subject may be asked for any kind of incident which produced favorable or unfavorable feelings. This technique was adapted to questionnaire form as shown in Appendix B. The questionnaire took the form of a 7x8½ inch booklet of seven pages. This critical incident questionnaire was administered in four classes at the University of Tennessee, and descriptive phrases were gathered and included in the item pool. A number of new items resulted from this questionnaire, but not enough new phrases were secured to warrant further administration of the form.

The item pool now contained about 1,000 items, and it took excessive reading to find a new idea in the records. A tentative conclusion reached at this point in the research was that extensiveness or size of sample was not a substitute for more intensive work with each subject student. However, depth interviewing may lead to concepts not easily communicated to raters by means of short phrases. Although there is no rigid necessity for short phrases as such in rating generally, there is need for behavioral referents; and so interviewing designed to get at "incidents" appears desirable. Interviewing was not attempted in any systematic fashion, however, because it is, if properly done, an expensive and time consuming process.

Figure 1 provides the distribution of ratings on the seven-point scale found on Page 4 of the last questionnaire (Appendix B). This distribution of teaching effectiveness of the "best-known" teacher is approximately symetrical. It was, therefore, considered feasible to construct the QCL so that the student would describe the "best-known" teacher. It was suspected that the "best-known" teacher might also be the "best" in most cases; if this were true, then little data would be obtained on the poor teachers. This actually turned out to be the case when the first QCL form was subsequently administered.

Editing and Selecting the QCL Phrases Used

A pool of some 1,000 phrases or words had been assembled by using the procedures described, and had been filed under the categories mentioned. It was proposed originally to build separate check lists

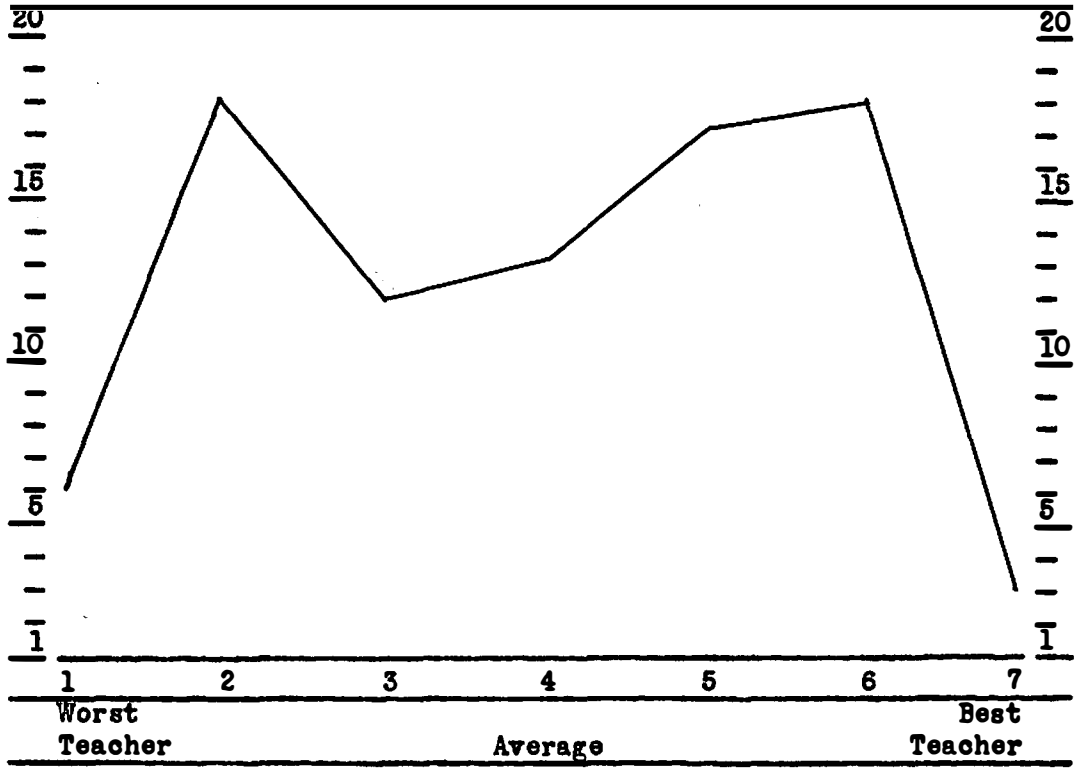


Figure 1. Distribution on seven-point scale, overall teaching of best-known teacher. Critical Incident Method, Revised. February, 1950. Number = 100.

for each of these categories. For example, one part of the check list would be concerned with lecturing, another with grading, etc. However, a rough screening eliminated only about 100 phrases, and it was estimated that continued screening would do little more than halve the pool.

The problem, then, was to administer a check list of perhaps 450 items, without excessive rater fatigue effects, to a representative sampling of college students. Classroom administration was considered the only feasible method. If administration were on an individual volunteer basis, various biases of interest, or other unknown systematic effects, would affect the results. This decision meant a maximum administration time of fifty minutes. However, rater fatigue allowance reduced the number of items possible still further. Separation of the check list into parts was therefore sacrificed in order to make the six forms of the QCL roughly comparable with respect to each of the categories. It should be noted that the final scale can be <u>scored</u> into part-scales.

The 900 items remaining after the first screening were reduced to 500 by further editing by the writer and were transferred from cards to form a list of items. "Judges" were employed to edit the items for meaningfulness, ratability, ambiguity, and duplication. Drs. Edward E. Cureton and Louise Witmer Cureton were the major judges. Various faculty and graduate student friends also checked the list of phrases.

As a result of this procedure some 400 items remained in the card pool of items. Since International Business Machine procedures were to be used in the analysis of results, IBM card structure determined, in part, the number of items that could be included on any one form of the

check list. The QCL forms as they stand represent a compromise between the amount of data desirable in the definition of the population (such as age, sex, academic year, etc. and the number of forms of the check list desired. For example, after minimum classification or definition data was determined, sixty-six spaces were left free on the IBM card. The number of phrases in the item pool divided by sixty-six resulted in six forms of the QCL. A few excess items were eliminated on the basis of criteria of ratability, overlapping with other items, etc. Use of more than six forms would have made it difficult to equate the forms for the categories mentioned earlier. Since ratings of teacher effectiveness follow the check list items, it was possible that overloading on a form on one category (for example, grading,) might result in a distribution of teacher effectiveness rating on this form markedly different from the other forms of the QCL. The number of items, sixty-six, fitted nicely the problems of both rater fatigue and administration time. Figuring on the basis of an examination consisting of 66 true-false items meant a predicted administration time of thirty minutes. A second, but important, consideration was that of teacher cooperation in administering the check lists. Half of a class period appeared to be the maximum contribution that could reasonably be expected of college teachers.

Therefore, six forms of the QCL of sixty-six items each were desired. To build them, the card pool was formed into six piles in such fashion that approximately equal numbers of cards from each category were used. Cards were then shuffled and laid out in random order. Each list was inspected to avoid conjunction of two phrases likely to affect each

other, for example, "respects opinions of students" and "courteous."

Qualification Check List Pilot Studies

Two pilot study editions of the QCL were used to study various issues of format and directions and the distribution of teachers rated on the "effectiveness" criterion scale.

The check lists were to be administered to classes throughout the country under widely different conditions. The administrator, if directions were followed, should be a graduate student with training in psychology. But the QCL might be given by an instructor with little training in education or psychology. Therefore, the check lists should be almost completely self-administering in order to avoid errors of administration. Directions and format thus should be checked against student response for errors of interpretation.

The heading, as printed, was designed to lend the QCL prestige and hence the serious consideration of those reading it.

The statement that the administrator cannot answer questions, as found under Orientation, resulted from experience in administering the first pilot study QCL. Students, when told to read the directions again, had little trouble, but attempts to explain seriously lengthened the administration time and may have been a source of bias.

All classification data was grouped in later editions of the QCL to avoid inteference with directions or carrying them out, and definitions of the letter ratings were simplified.

The "effectiveness-as-a-college-teacher" and "liking-as-a-person"

nine-point criterion scales (herein referred to as "effectiveness" and "person" scales) were changed in the second edition of the QCL to prevent students from checking more than one rating. The criterion scales were separated spatially in the printed form, and the "effectiveness" scale was reversed to prevent automatic checking. Very few of the printed QCL's in the final administration had more than one box checked.

Kind of teaching in Classification Data was inadequate in the first two editions, whereas the ranking method gave little difficulty in the printed QCL.

The first edition of the QCL asked the student to "select the teacher you know best." This followed the practice of the critical incident questionnaire used to collect phrases for the item pool. Figure 2, however, reveals a very different distribution of effectiveness rating on the first pilot edition of the QCL. This distribution indicates that the "ideal" method of having the student select a teacher, describe him on the check list, and then rate him, was not feasible. Such skewed distributions would result in very few poor teachers.

The questionnaire and pilot QCL form were compared to see why the distributions differed. The instruction page of the critical incident questionnaire contained the word critical and it was underlined; it paved the way generally in hedonic tone for unpleasant matters, and this set apparently carried over to selecting the "best known" teacher as required on the next page of the questionnaire. With this set, the teacher was rated poor about half the time, even though an incident favorable to the teacher intervened on Page 3. Therefore, the effect

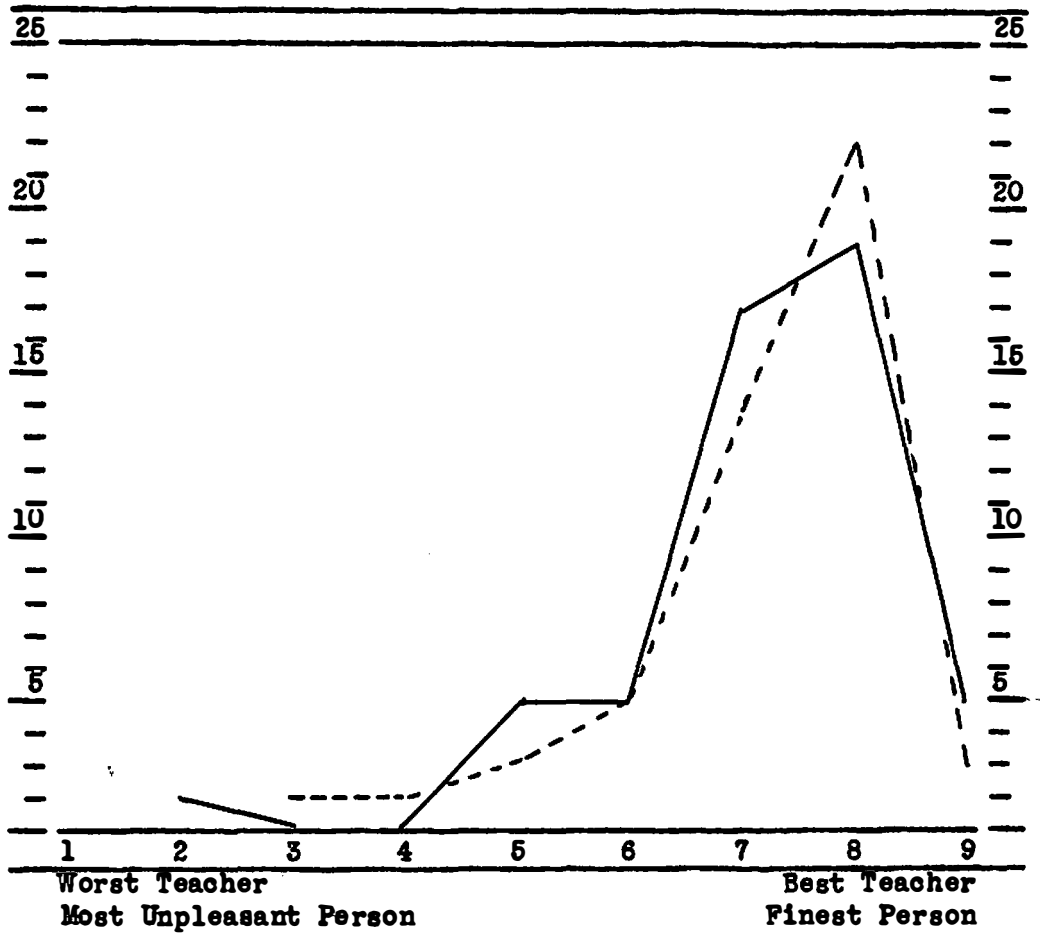


Figure 2. Distribution on nine-point scale, first edition of pilot Qualification Check List. _____ = rating of teacher effectiveness. ------- = rating of teacher as a person.

of such an approach was tested in the second pilot edition of the QCL to avoid possible forcing of the distribution too far, and to promote the use of any "critical incident" as a sort of stimulus to student recall of a particular class other than the one of administration. The fact that a teacher could handle a critical situation well was stressed by underlining in the directions. It was felt that if a vivid memory of a particular teacher could be recalled it would serve as a fairly stable nucleus around which additional recall could cluster. The "cards were stacked" in favor of poor teacher recall by placing examples of poor handling of incidents first, and also by underlining "critical."

In studying the first pilot edition of the QCL further, it was noticed that all of the phrases were stated positively. The cumulative effect of sixty-six such positive statements on the effectiveness and person ratings may be substantial in forcing it toward the high end. Furthermore, consideration of the purpose or meaning of each phrase is involved. We are asking a student if a teacher "keeps promises," for example. It seemed reasonable to indicate this, so a question mark was placed after each phrase. Putting each phrase in question form may improve the distribution of rating each phrase, since a change in set is less likely to occur as the student goes through the list.

The results achieved by these changes in the second pilot edition of the QCL are shown in Figures 3 and 4. Skewness is still present. The difference in the Georgia Tech and Tennessee distributions led to the belief that the national sampling might be less skewed. The fact that the distribution still turned out skewed in the final QCL administration is

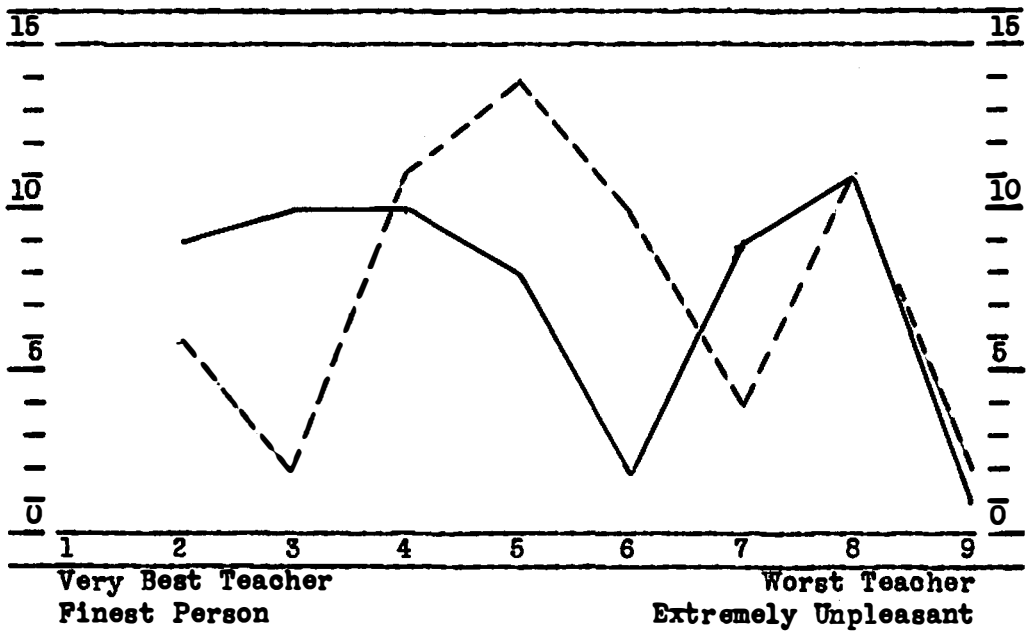


Figure 3. Distribution on nine-point scale, second pilot study of Qualification Check List at Georgia Institute of Technology. _____ = rating of teacher effectiveness. ------ = rating of teacher as a person.

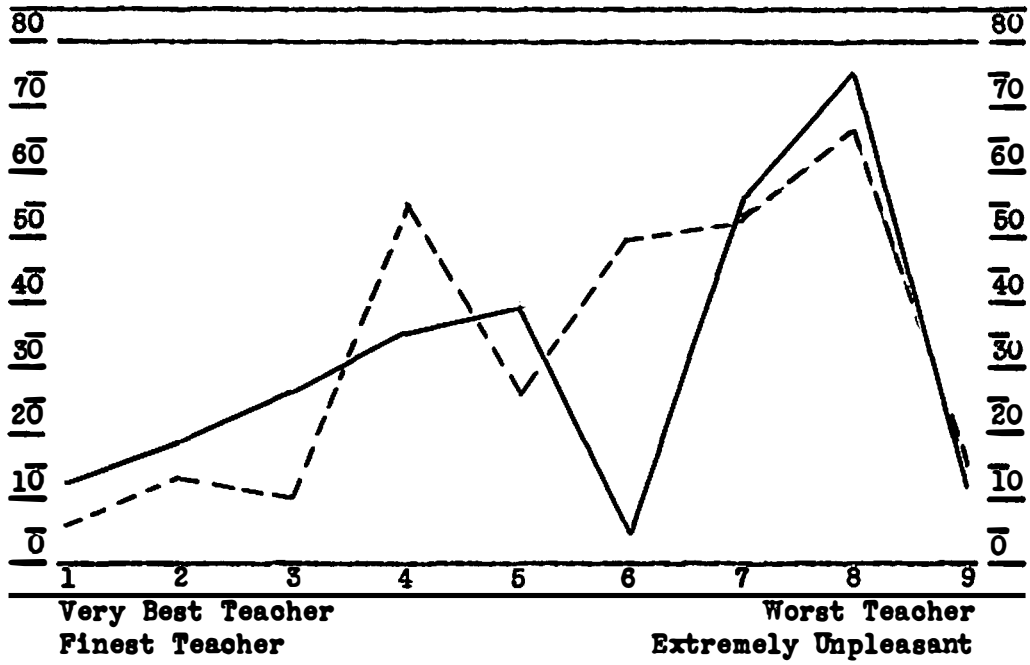


Figure 4. Distribution on nine-point scale, second pilot study of Qualification Check List at University of Tennessee. _____ = rating of teacher effectiveness. ------ = rating of teacher as a person.

not evidence against these concepts. It may be that the very high frequency on eight, "one of the two or three best teachers I have known," as compared to that on six and seven in the final result is a function of the phrases used. Johnson (12) found that various positions of a rating scale, regardless of what is being rated, are not used with equal frequency. Certain positions have a special attractiveness. Perhaps without modification of the early QCL editions, the distribution would have been even more skewed. The advantage of avoiding a "good" or "bad" stereotype, together with use of all of a distribution rather than its ends, is considered worth the disadvantage of a skewed distribution.

Statistical Results in the Pilot Study

The second pilot study edition of the QCL was also used to check on discrimination indices and the correlation between teaching effectiveness and liking as a person ratings.

Forced-choice methodology requires a number of items of high discrimination and some of moderate discrimination so that in pairing items sufficiently large differences can be achieved. Since all the words or phrases were positive, and since they had been rather heavily edited, it was necessary to check to see if a sufficient number of moderate discrimination items existed.

Form G of the QCL was used. "High" and "low" groups were selected by using the highest and lowest three ratings on the nine-point effectiveness scale. The mean of the high group was compared with the mean of the

low group for each item. Very Good (VG) was coded four; Very Poor (VP), zero. The difference in means for items ranged from .1 to 2.1. A difference of .1 had a t of .37 and was not significant. A difference in means of .5 had a t of 1.82 (p = .03), and a difference of .8 had a t of 8.

Considering the size of the sample contemplated for the printed QCL (1,000) and the fact that there would be 395 items, it was considered advisable to go ahead even though there was an excess of highly discriminating items, some of moderate discrimination, and few items low in discrimination.

The product-moment correlation between the ratings on the nine-point effectiveness and person scales was .73. The 1 per cent confidence limits, using Z' were .58 and .76. Therefore, since the interscale correlation was substantially lower than unity, correlation of each item with both effectiveness and person ratings would be justified. In other words, the effectiveness and person ratings were getting at something different, and therefore item relationships should be different.

Preparation of the Printed QCL

The major considerations in constructing the printed QCL were cost, weight, attractiveness, legibility, and ease of transferring data to IBM cards. Format considerations were mentioned earlier. Mr. Thomas Greene, Head of the Central Editorial Service, University of Tennessee Public Relations Office, was very helpful with suggestions as to paper size and design. A photo-offset process was used. With plates of this size,

camera distortion may be a problem, depending on the equipment used. The forms were also printed by Central Editorial Service, University of Tennessee. Form M of the Qualification Check List as printed and used is shown in Appendix C. The item content of the other five forms may be obtained from Appendix F.

CHAPTER V

ADMINISTRATION OF THE QUALIFICATION CHECK LIST

Method of Obtaining the Sample

Members of the staff of the Department of Philosophy and Psychology, University of Tennessee, provided a list of names of their colleagues in psychology throughout the country. Form letters requesting administration of the Check List were prepared in ribbon copy and signed by the staff member who contributed the name. Returns on these letters were heavy in psychology and education classes. It was perhaps unrealistic to expect the recipients to go outside of their departments when, as many indicated, the Qualification Check Lists had instructional value in their own classes! This fact was not fully realized until members of the American Educational Research Association, canvassed by personal letter from the writer, responded in like fashion.

Because of a small return from the methods described, college catalogs were selected at random from library shelves, names of deans, chairmen, heads of departments, and professors were chosen at random, and personal letters were sent to two or three at each college. This method borders on the "accidental" rather than random. However, a more systematic method would have been aborted by the heavier acceptances by those teachers sympathetic to research in this area. It is assumed that there is no significant relationship between this sympathy and student response, except that more "good" teachers were selected than might otherwise be the case.

Because of the already heavy returns in psychology and education, staff in these subjects were not solicited. An attempt was made to choose, within any one college catalog, teachers of "shop," "field," and laboratory subjects.

During the first week of February, 1952, the last of the QCL forms were received from the printer. They were then checked for legibility and the six forms, H through M, were arranged in cycles. Three hundred and fifty-two letters of request had been mailed; 42 per cent replied.

During the next few months, until the last batch of forms was mailed on April 29, 1952, various supplementary letters of explanation and follow up were sent out. As a result of letters of all kinds, there were one hundred and eleven acceptances; this constitutes a 31 per cent return.

Return of the completed QCL's was 90 per cent of those shipped. Two mailings were "lost," six were returned after the deadline for punching the IBM cards, and three mailings were not returned at all. This meant that ninety-two of the one hundred and three shipments were usable. Therefore, three hundred and fifty-two requests resulted in ninety-two schools, or 26 per cent, being punched on IBM cards. This percentage would have been higher had it not been necessary to refuse eight acceptances because of prior overloading in psychology and education.

Stamped, return-addressed envelopes accompanied all correspondence. Packages were weighed and stamps and envelopes for their return were sent

with the QCL blanks. Prompt response was made to all acceptances. Postcard reminders were used to follow up when the completed forms were not returned within five weeks.

The original method of administering the QCL, using psychology staff throughout the country to act as selectors and solicitors of classes and teachers on their respective campuses was ineffective. The three enclosures, as shown in Appendix D, sent with the blanks were therefore modified. When the new method of direct mail request was adopted, the Directions-for-Administration page was changed by circling the x-ing out First steps in order to avoid confusion. The letter intended as an aid to the psychology teacher of graduate student in securing classes (2nd page of Appendix D) was eliminated. Reference was made to both Directions and Administration Log in the covering letter when the blanks were shipped.

Analysis of the Schools Used

As the returns were received, the item read first, because of its significance for the whole research, was the statement of response tone on the Administration Log. Pilot studies had been reassuring, but administration under conditions which demanded more of the QCL might be less so. Only two of the one hundred and twenty-four classes of the sample had a response tone interpreted as unfavorable to the research by the administrator. Apparently the response justifies confidence that the students followed instructions. Some of the administrators did not; some apparently administered the QCL's themselves, and a few did not

complete or return the Administration Log.

Scanning of the individual QCL's prior to IBM punching reinforced this conclusion. Very few double ratings on effectiveness or person ratings were noted. Few forms were discarded because of failure to complete each item under Classification Data. Only one pattern response was noted; i.e. 12345, 12345, etc.

A good deal of white space was deliberately provided under Comments in the QCL and no cues to its use were included, in the belief that it would tap attitudes toward the QCL. Comparatively little use of it was made; those comments made dealt almost exclusively with the teacher described rather than the QCL. There were a number of favorable comments on the QCL and only one or two unfavorable ones.

The geographical distribution of the QCL sample is shown in Appendix E. A rough correspondence with population density may be noted; however, the West and Northwest seem to lack representation. It would seem that the results shown are adequate in case geography is significant. Additional schools could be added to the sample later if further research indicates it is necessary.

The schools were largely coeducational. Five women's and five men's colleges are included in the sample. They can be used in later research on sex differences. In the larger schools, there are very few purely male or female institutions, and even in the smaller ones, exceptions are made for some classes or for veterans.

The size and level of the schools in the sample is indicated in Tables I and II. The data are rough, since they are 1946 figures

TABLE I

DISTRIBUTION OF SCHOOL SIZE IN QUALIFICATION CHECK LIST IN TERMS OF STUDENT ENROLLMENT

School Enrollment to Nearest 1,000	Number of Schools
Below 1,000 students	9
1,000 students	8
2,000 students	4
3-5,000 students	12
Over 6,000 students	37
Total	70

TABLE II

LEVEL OF SCHOOLS USED IN QUALIFICATION CHECK LIST SAMPLE ACCORDING TO DEGREE CONFERRED

Degree	Number of Schools
A.B.	13
M.A.	22
Ph.D.	35
Total	70

and "level" probably is not accurately indicated by "Ph.D." However, there is sufficient range in the sample. As shown in Table I, perhaps there are too many large universities and not enough medium-sized schools. If later research shows significant practical differences between the smallest and largest schools, then the forced-choice scale could be restricted to one or the other or corrections made. The sample should provide enough cases for both large and small schools; if not, supplementary data may be obtained.

Table II shows that in terms of level, the sample is weighted with schools having doctoral programs. This may be misleading since a school was placed in this category if only one doctoral degree was awarded. If later comparisons justify it, more accurate classification methods can be employed. Perhaps a composite criterion of both size and level would be the best to use in further research.

Analysis of the Classes Used

Since provision was made under Classification Data on the QCL for student age and school year, it was considered unnecessary to describe the class in which the QCL was administered as a sophomore course, junior class, etc. Freshman classes were eliminated at the outset in the Directions for Administration, and any QCL's marked other than sophomore, junior, or senior were not included in the sample.

The names of the courses or classes in which the QCL's were administered are not necessarily descriptive of the teacher rated. In fact, specific instructions were given to the student not to describe

the instructor of the class, since this would limit the size of the sample of teachers described very sharply. The teacher described and rated is thus unknown.

The fact of anonymity meant that no control, in any strict sense, could be exercised over the sample of teachers rated. There could be no breakdown of teachers rated by name, subject taught, etc.

An anonymous method was considered necessary, not only because it would be easier to find teachers willing to have it administered in their classes, but because it was necessary to dispel student fear and hostility. Observation of response to the check list of both teacher and student during pilot-study administration seemed to justify this decision. It is true, that after administration some students commented that it was "too bad a certain teacher couldn't see it," etc. Prior to administration a heightened attention suggested some tension.

The most important reason, however, for accepting an "unknown" sample is one of controlling bias. Honesty of response was crucial, since any leniency, through fear of results being made known (or its converse "apple polishing,") or any attempts at revenge, or any similar motive, would distort the discrimination indices obtained. In essence, what was needed was a group of "disinterested judges" to decide which words or phrases were descriptive of good teaching, much as has been done in constructing check-list scales in industry. If students are to be used as judges, it is then quite clear that nothing should interfere with their decisions. The average rating could not be used,

as in the cross validation scale, so that bias becomes important.

Perhaps in a loose sense some control might have been exercised because of the fact that a biology major, for example, might be more likely to describe a biology instructor than any other kind. Therefore, early overloading in psychology and education was of some concern, even though it is true that psychology undergraduates are heterogeneous rather than homogeneous in curriculum.

Table III shows the results in terms of major curricula, even though it is not safe to infer that a distribution of teachers actually rated would be parallel, in view of the limitations just described. All "education" classes are grouped together, regardless of level. Here again, provision for further research has been made through "kind of teaching" under Classification Data on the QCL. The sampling presented in Table III should provide enough of these various teaching situations or methods so that later research will be able to isolate differences, if any.

Therefore, it is difficult to estimate the effectiveness of the sampling. When the Classification Data are analyzed, it is theoretically possible to compare them with estimates of the student population from other sources. It is now regarded as a refinement unnecessary to the research objectives accepted.

The administration time, provided on each Administration Log, corresponded to that of pilot study QCL's. The range was from fifteen to thirty-five minutes. By inspection, the mode was about twenty-five minutes. Administration time is significant as an indication that the

TABLE III

DISTRIBUTION OF THE QUALIFICATION CHECK LIST SAMPLE BY TYPE OF CLASS IN WHICH FORMS WERE ADMINISTERED

Type of Class	Number of Classes
Agrioulture	3
Biology	6
Business	7
Chemistry	2
Economics	2
Education	25
Engineering	9
English	4
Home Economics	2
Industrial Arts	1
Journalism	3
Language	2
Law	1
Mathematics	3
Philosophy	4
Psychology	35
Sociology	17
Total	124

administration proceeded according to plan. It was also useful later in estimating the administration of the validation scale.

The Administration Log data on day of week and hour of day can also be used in further research to investigate the effect of these variables on effectiveness and person ratings.

CHAPTER VI

ANALYSIS OF THE QUALIFICATION CHECK LIST DATA

Mechanics of the Analysis

As the completed QCL's were received from the schools, they were coded and scanned for compliance with instructions and for completeness. Information from the QCL's remaining after this screening was then punched on International Business Machines standard cards.

The card punching was done by the writer's wife and trainees in the International Business Machines office in Montgomery, Alabama. The cards were not verified. It is assumed that errors were random. Spot inspections showed very few errors; four cards of the 3,640 used in the preliminary run were mispunched, which resulted in a final sample of 3,636 cases. This was considered very good punching by the civilian chief of statistical services, Maxwell Air Force Base.

Correspondence with the Secretary, Air University, United States Air Force, produced permission to use the facilities of Statistical Services Section of the Command on a non-priority basis. While the cards were being punched, consultations with Mr. James Reynolds, at that time project director for Statistical Services, resulted in the methods described in this section.

Although the distributions of the nine-point ratings for person and effectiveness scales were then unknown, it was decided that product-moment correlations between them and the five-point item scores should

be used. Inspection of other possible methods showed that most of them sacrifice data. For example, use of the difference in mean of high and low groups as a discrimination index sacrifices the middle group. Various other methods were either cumbersome, or made assumptions as to normality of one of the variables anyway, so that product-moment method might as well be used. The distributions did turn out skewed, but most of the regressions were close to linear. The IBM methods used yielded a 5x9 scattergram, as well as the necessary sums, sums of squares, and sum of products, for each item-scale correlation, for both criterion scales. The final computations were done on a desk calculator.

Results of the Qualification Check List Analysis

As shown in Figures 5 and 6, the scale distributions are skewed, but not to such a degree that the results are unusable. It is assumed that students found some descriptive categories in the criterion scales quite close together in meaning. After locating a teacher generally on the scale, they appear to have marked the most appealing or acceptable of two adjacent statements.

Figures 7 and 8 show the extreme frequency surfaces for each of the six forms of the effectiveness and person scales. At each of the nine rating points in these figures, the highest frequency for any of the QCL forms is shown as well as the lowest. Thus the distribution of each of the six QCL forms, when plotted individually, falls within these upper and lower limits. The range from high to low at each scale

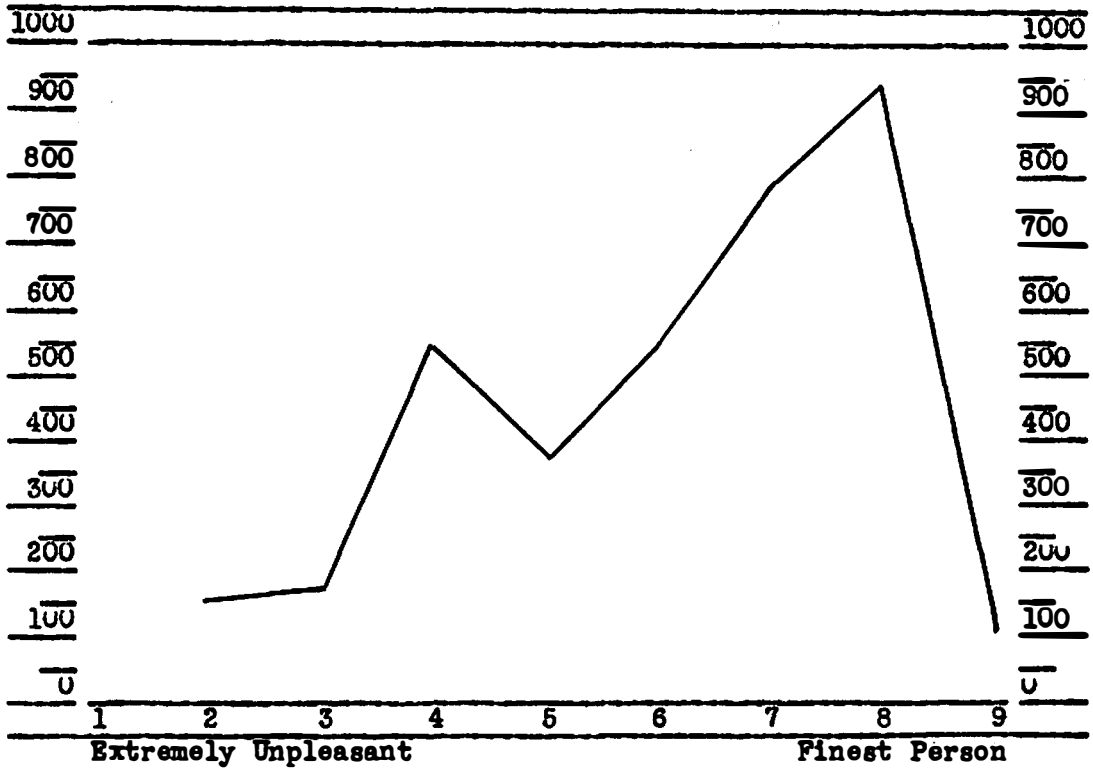


Figure 5. Distribution on nine-point scale of ratings on "liking as a person" for all forms of the Qualification Check List.

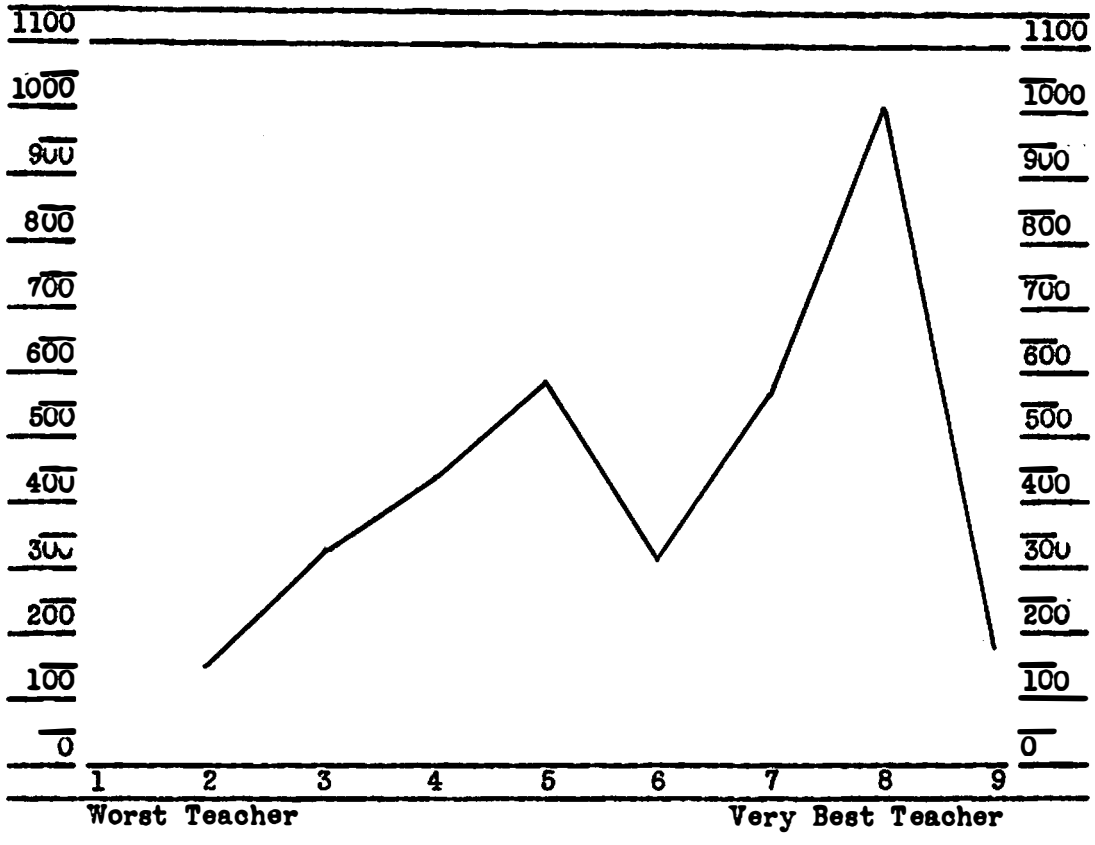


Figure 6. Distribution on nine-point scale of ratings on "effectiveness as a teacher" for all forms of the Qualification Check List.

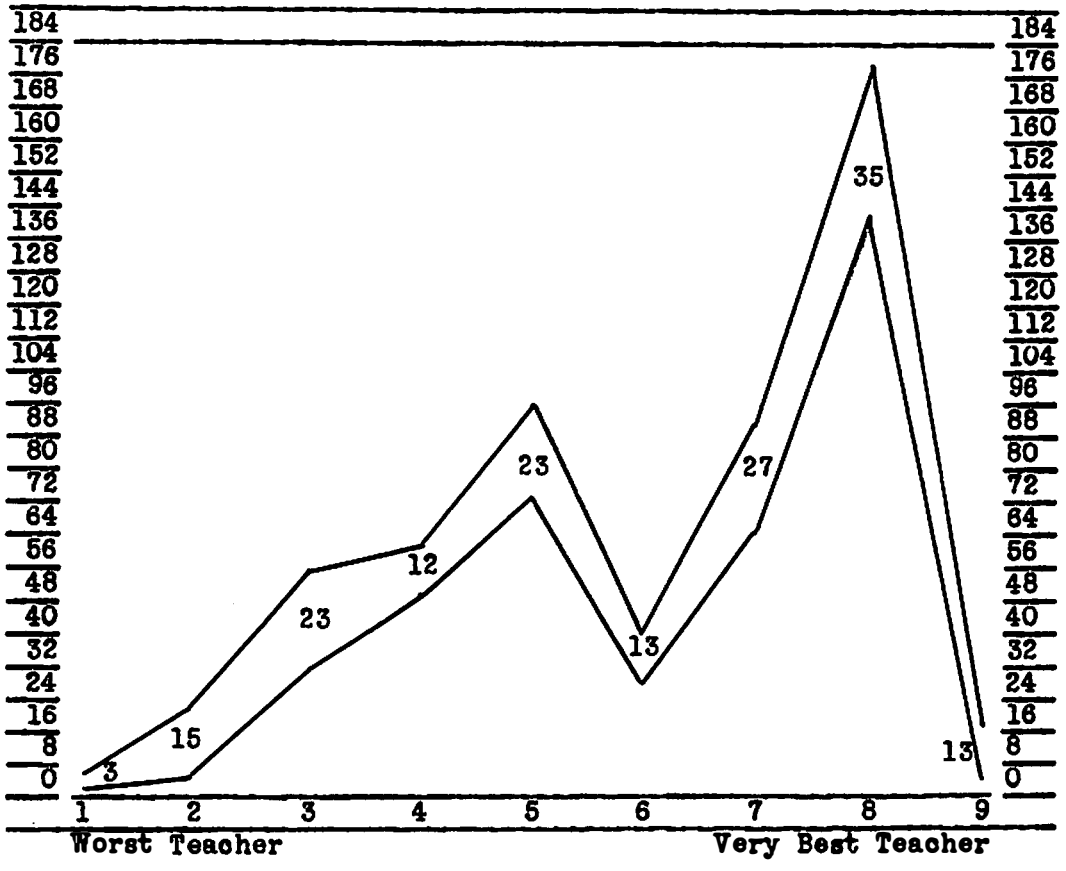


Figure 7. Maximum and minimum frequency among the six Qualification Check List forms for each scale point in "effectiveness as a teacher." Numbers show range of highest and lowest ratings at each scale point.

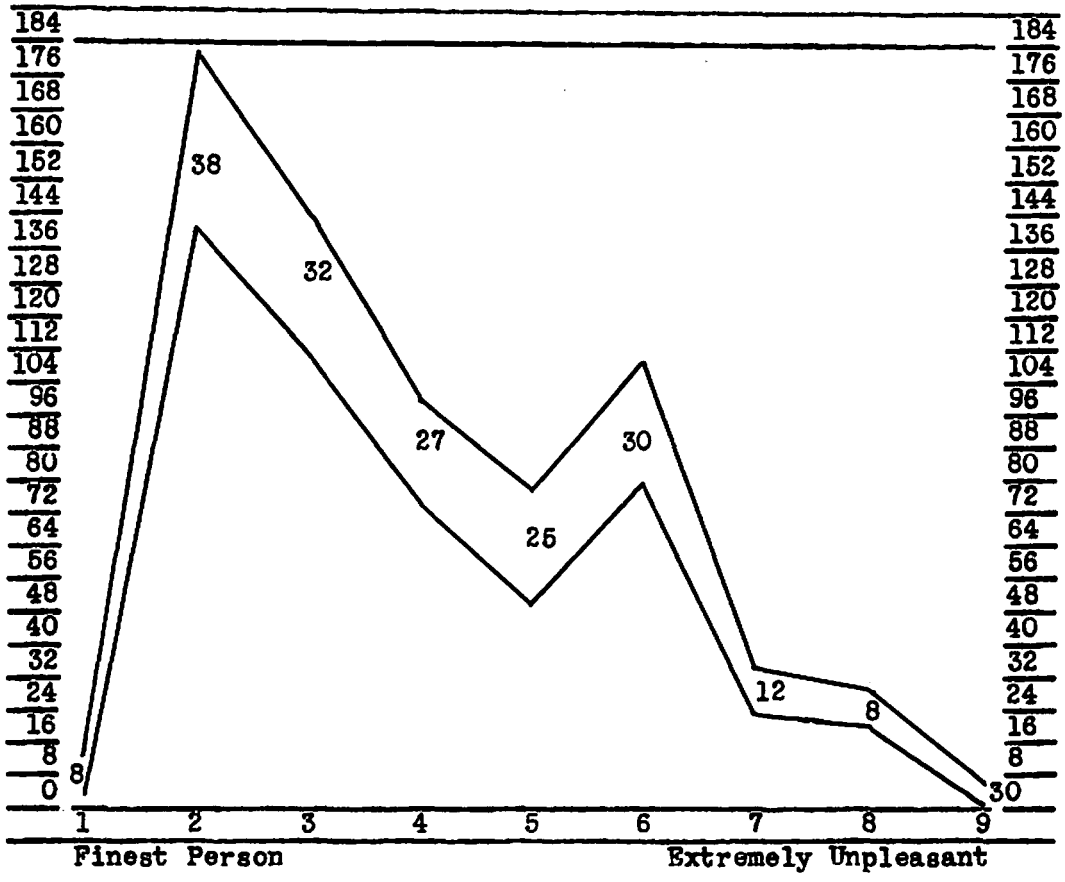


Figure 8. Maximum and minimum frequency among the six Qualification Check List forms for each scale point in "liking as a person." Numbers show range of highest and lowest ratings at each scale point.

point among the six forms is shown. It will be seen that the distribution of each of the forms is about the same. Since the student first read the sixty-six phrases on the QCL and then rated the teacher on the two nine-point scales, and since the distributions are similar, it would appear that either the QCL phrases had no effect on the nine-point ratings or that they had an equal effect upon the two types of ratings. Therefore, the six forms of the QCL were considered comparable, and form differences were disregarded in building the validation scale.

Ratings from the nine-point effectiveness and person scales were again correlated. Whereas the pilot study QCL produced a product moment correlation of .73, the QCL as administered yielded a correlation of .54.

Turning to the three hundred and ninety-four items of the QCL, a study of the mean rating for each was made. Where 1 stands for a "very poor" rating and 5 indicates a "very good" rating, the range in mean ratings was from 1.47 to 3.98. The frequency distribution of means is shown in Figure 9.

Figure 10 shows the frequency distribution of correlations between each item and the effectiveness and person ratings. It should be noted that the correlations are expressed as Pearson correlations; Fisher's Z' transformation has not been used. This figure should be contrasted with the sampling distribution of correlation as found in textbooks on statistics. All item-scale correlations were positive.

A 5x8-inch card was prepared for each of the three-hundred and ninety-four items. Entries on these cards were made for the item mean, the item-effectiveness correlation (r_{ie}), and the item-person

```
Magnitude
  of
Means

3.90 to 3.99   x 1
3.80 to 3.89   x 1
3.70 to 3.79
3.60 to 3.69
3.50 to 3.59
3.40 to 3.49   x 1
3.30 to 3.39   xxx 3
3.20 to 3.29   x 1
3.10 to 3.19   x 1
3.00 to 3.09   xxxxx 5
2.90 to 2.99   xxxxxxxx 8
2.80 to 2.89   xxxxxxxxxxxxx 13
2.70 to 2.79   xxxxxxxxxxxxxxxxxxxxxxxxxx 26
2.60 to 2.69   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx 32
2.50 to 2.59   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx 31
2.40 to 2.49   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx 52
2.30 to 2.39   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx 57
2.20 to 2.29   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx 43
2.10 to 2.19   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx 31
2.00 to 2.09   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx 37
1.90 to 1.99   xxxxxxxxxxxxxxxxxxxx 20
1.80 to 1.89   xxxxxxxxxxxxxxxxx 17
1.70 to 1.79   xxxxx 5
1.60 to 1.69   xxxxx 5
1.50 to 1.59   xx 2
1.40 to 1.49   x 1

                         Number of Means (total 394)
```

Figure 9. Frequency distribution of item means for 394 items of the Qualification Check List. (1 is low, 5 is high.)

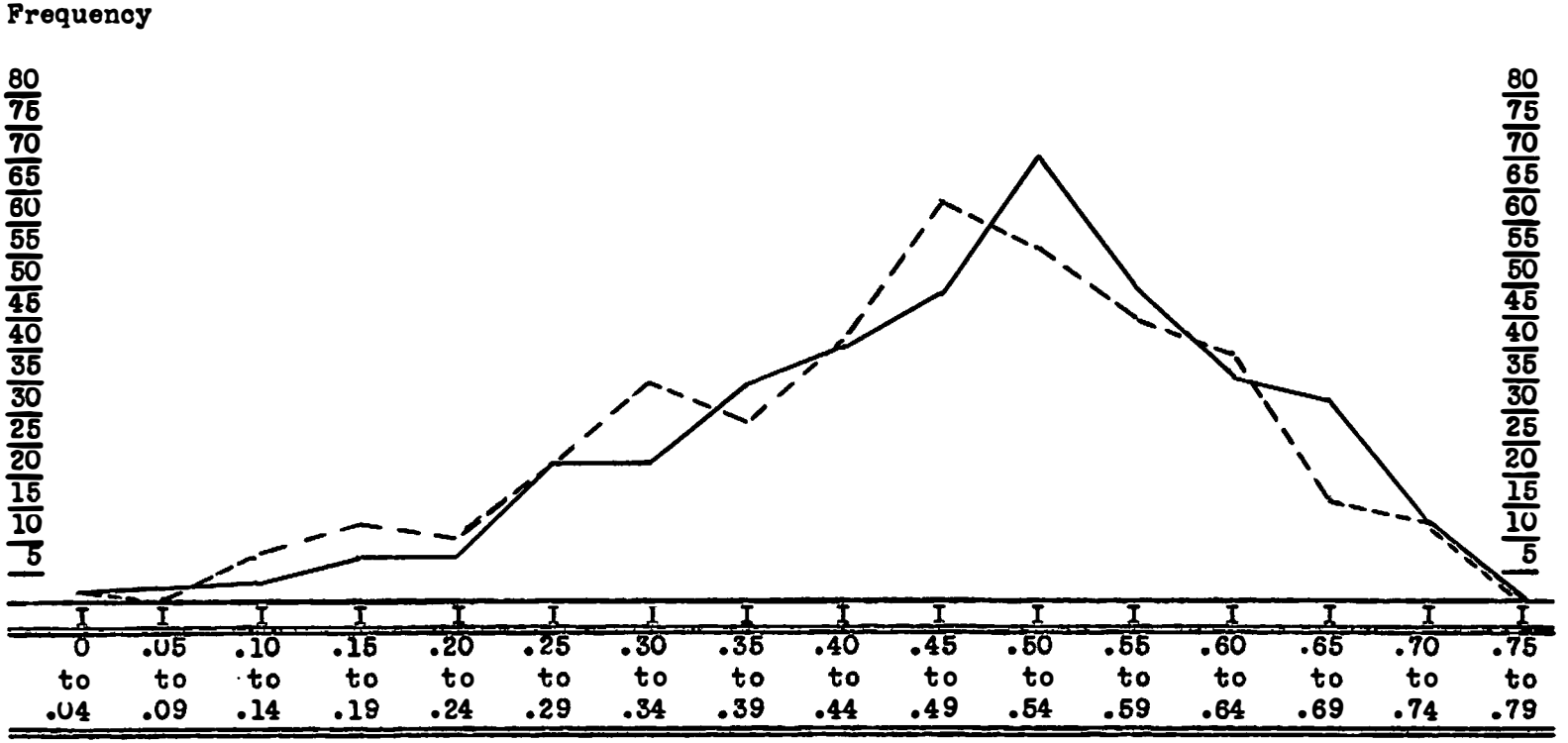


Figure 10. Frequency distribution of correlations of each item of the Qualification Check List with the nine-point "effectiveness as a teacher" and "liking as a person" scales.
_____ = "effectiveness," ------ = "person."

correlation (r_{ip}). The results are presented in Appendix F.

CHAPTER VII

THE CONSTRUCTION AND ADMINISTRATION OF THE VALIDATION SCALE

Introduction

From the analysis of the Qualification Check List data, the mean of each of the three hundred and ninety-four items on the five-point scale and the correlations between each item and the person criterion scale (r_{ip}) and the effectiveness criterion scale (r_{ie}) had been obtained. From previous discussion it will be recalled that it had been decided to proceed on the basis of the minimum number of assumptions; in this case, to cross validate with another sample by item rather than by couplet. The scale would then appear as a graphic scale. However, in order to control on possible item-position effects in future forced-choice administrations of the scale, and to permit scoring as if forced-choice, couplets had to be built according to forced-choice rationale. It will be recalled that the mean had been decided upon as a characteristicness index, the r_{ip} statistic as a biasability index, and the r_{ie} statistic as a discrimination index. The major problem in constructing the validation scale was, then, to build each couplet with as large a difference in discrimination as possible, and at the same time maintain equivalence of items on both the characteristicness and biasability statistics.

Rationale For The Couplets

The rationale of the indices must be reviewed briefly to explain the procedure used in building the couplets. Equal item means signifies that raters, in general, are equally likely to make the two statements about the person rated. When items are equal in correlation with the person scale, a check mark on one of the items predicts the rater's liking for the person rated just as well as does a check mark on the other item.

When the biasability indexes are equal, but the characteristicness indexes are not, raters generally are more likely to check the item with the higher mean.

In the case of equal characteristicness, but different biasability indexes, raters are equally likely to check the two items, except that a more acceptable person will be checked on the item with the higher r_{ip}. With items unequal on r_{ip} the student would bias the rating by marking the item with higher r_{ip}.

At the time the scale was constructed, no way was seen to combine these two indexes into an index of acceptability, inasmuch as an index of mean bias tendency was not available. Therefore, a graphic technique was designed to control both the characteristicness and the biasability index while building couplets with as large differences in item discrimination as possible.

Procedure in Building Couplets

Following a suggestion from Dr. E. E. Cureton, a bivariate chart was prepared on which item mean and item r_{ip} were employed to fix the position of each of the three hundred and ninety-four items. The ratio of the mean standard error of the Z' transformation of the r_{ip} to the mean standard error of the mean of the items was used to scale the distances. The mean standard error of the mean was computed, independently by two different methods. One method used modal frequencies on the various forms of the QCL as a sampling method, using the five-point scale values. Another method took a small sample of means from each of the forms. The mean standard error of the mean was consistently equal to .005 (rounded) by all methods. Using an N of 560 as typical of the size of the student sample for any form of the QCL, the mean standard error of Z' was .013 (rounded). The ratio of the mean standard error of the mean to the mean standard error of Z' was 2.7 (rounded). Using the 5 per cent significance level (1.96 sigma) of the standard error of the Z' difference as a guide ($\simeq$ .05), the chart was scanned to pick up item pairs close together on the dimensions of mean difference and r_{ip} difference. These items were located by referring the QCL form letter and item numbers back to the item cards to obtain the item correlations with the effectiveness criterion scale. Using a worksheet, the r_{ie} item differences were computed. If the criterion of 20 Z' points difference or greater was met, the pair or couplet was used.

In this fashion, fifty-eight couplets were constructed. Thus

one hundred and sixteen items of the original three hundred and ninety-four were used. This is 29 per cent usage, or about 70 per cent shrinkage on an arithmetical basis. The loss in terms of items of diagnostic value in improving quality of instruction is unknown.

Figure 11 contrasts the distribution of the r_{ie} Z' for the group of highly discriminating items with that of the lower group. It will be noted that, in general, the couplets are composed of items of high versus moderate discriminating power. In view of the pile up on the high side of the criterion scales, as shown in Figure 7, Chapter VI, it would seem that the crucial rating problem is that of discriminating among teachers who are average or better. Therefore, the fact that the couplets turned out as shown was regarded as very encouraging in that they would permit relatively fine discriminations.

Scale Construction

The scale itself was constructed with a minimum or common denominator set of directions so as to facilitate its use in a variety of situations. The meaning of the letter ratings was simplified as compared to the QCL set, and the scale values were increased from 5 to 15 to allow discrimination between Poor to Fair, Fair to Good, and Good to Outstanding, as shown on the scale in Appendix G. This last procedure was impossible in the QCL because of format problems. The couplets were first organized so that the highly discriminating item came first, then a number of these were reversed; so that the scale, when administered in forced-choice form, would not have this "key" item consistently first.

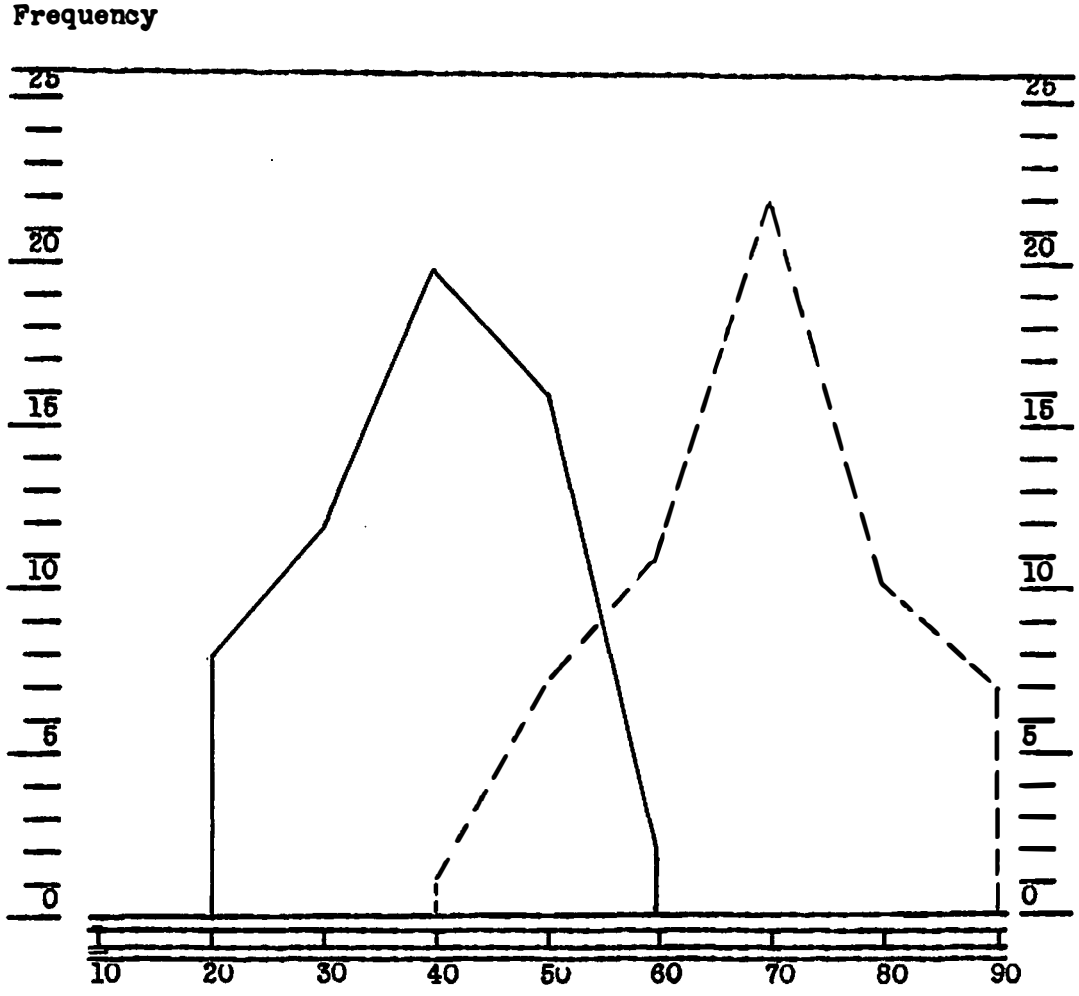


Figure 11. Distributions of discrimination index magnitudes in couplet construction. ——— = low discriminating items. ------ = high discriminating items.

There is, then, no item position effect to worry about in terms of systematic bias. The reasons for not using forced-choice format in constructing the scale have been discussed.

The criterion scales were identical with those of the QCL; the person scale came first, and the continuum for the effectiveness scale was reversed; therefore, comparisons with the cross-validation sample are possible.

The writer is indebted to Dr. E. E. Cureton and Mr. Thomas Greene, who not only made valuable suggestions, but used the writer's sketches to produce the printed form.

Administration of the Scale

In the fall of 1953, another "direct mail" campaign was begun to secure cooperation in administering the validation form of the scale. The method first used was to go through the correspondence accumulated on the QCL and request that the scale be administered by those who did not cooperate in the QCL administration. When this mine petered out, a new sample was solicited. Acceptance from both methods was so small that recourse was made to the QCL administrators. This step was taken reluctantly because of the desirability of a completely different and independent sample. However, in view of the fact that approximately fifteen months had elapsed and that students were describing all the members of a department in the validation scale, the assumption that these two samples are independent ones is regarded as reasonable. The

major source of difficulty in obtaining administration of the scale, aside from the fact that named teachers were rated, was that the scales were provided free of charge only if complete (or nearly so) departments cooperated. It took approximately six months of correspondence to obtain the nine departments and seventy-one teachers of the cross-validation sample, in spite of a promise to score the scale and interpret the results for those participating.

The Manual of Directions employed to obtain participation is shown in Appendix H. The Manual provided information on background and purpose, use of the scale, restrictions on use of the scale, instructions to students, details of administration, scoring and interpretation of results, costs, shipping instructions, ordering, and directions for local scoring. An Administration Log, completed upon administration and returned so as to provide information on instructions used and response tone of the students is attached to this appendix.

Various aspects of this procedure require attention. In agreement with the decision to avoid an either-or restriction, i.e. teacher or administrator use, the Manual made it clear that any kind of local arrangements would be satisfactory. In this connection, the Administration Log provides for identification by department of the use made of the scale. A very definite restriction, however, was mentioned twice in the Manual; namely, that some kind of practical use of the scale had to be explained to the student as the basis for requesting student rating of the instructor. In other words, the students could <u>not</u> be told that the scale was administered "for research purposes." The reason for this

is quite obvious; if conclusions are to be drawn and items cross-validated, the crucial test is one of determining the shrinkage of statistics produced by operational use of the scale.

The option of local scoring by the departments was exercised in very few cases. A check of tallying was made and as a result all such scales were retallied by the writer.

Before the validation work to be reported in the next chapter could begin, the promise made in the Manual to score the results and to interpret them had to be kept. The report sent to the departments, which contains the mean rating on each item for the group of seventy-one teachers, is shown in Appendix I. Even though this had to be somewhat of a stopgap measure, it required substantial machine time and further delayed work of greater relevance to the thesis.

CHAPTER VIII

VALIDATION OF THE SCALE

Introduction

After the scales had been tallied and the interim report to the participants, as shown in the previous chapter and discussed later in this chapter, had been prepared, the data was treated so as to obtain answers to questions germane to forced-choice methodology. Such problems as the variation of scores on the effectiveness criterion scale across departments (schools); the effect on item correlation of administering the scale with a "live" teacher rather than an unknown one as in the Qualification Check List; the validation of couplets, i.e. item difference scores against the criterion; the validities obtained from different scoring methods; and various related considerations are presented and discussed in this chapter. A major purpose underlying all of this was, of course, the construction of the forced-choice scale in its final form.

The Cross-Validation Sample

In view of the promise made in the letter of request to avoid comparisons by name, class, department, or college, a description of the cross-validation sample becomes difficult in view of the comparisons made or projected for future research. Table IV provides the number

TABLE IV

NUMBER AND KINDS OF CLASSES IN CROSS-VALIDATION SAMPLE

College	No. of Classes	Kind of Class	No. of Students
A	7	Journalism	187
B	4	Psychology	106
C	12	Agriculture	253
D	4	Business	114
E	20	Business	569
F	4	Engineering	66
G	3	Education	84
H	12	Business	439
I	4	Mathematics	88
9 Schools	71 Classes	9 Departments	1906 Students

and kinds of departments by college. The number of student raters for each department is also shown.

Procedure

The data were processed by tallying ratings, by teacher, on blank scales. Another set of blank scales was used as a worksheet to enter the sum of the scores by item, coding from 0 to 14, from Poor through Outstanding letter ratings. The number of students was then recorded and the mean rating for each of the one hundred and sixteen items computed. Using the formula, $N_1\bar{X}_1+N_2\bar{X}_2\ldots N_c\bar{X}_k$, the mean of means, first by departments and then for the total sample, was computed. This last statistic was used in reporting to the participating departments. The range, mean, and standard deviations for the two criterion scales were computed and reported in the same way. Thus the basic data in the following discussion consists of the teacher score on each item on the average over the whole class, i.e. the mean. Standard deviations have not yet been computed.

The Criterion Scales

Since the criterion scales were identical with those of the Qualification Check List, a comparison of the distributions resulting from the two administrations is of interest.

Figure 12 compares the per cent of the sample at each of the nine scale points for the effectiveness and person ratings when the class averages for the 71 teachers are used (A in the figure); the distribution

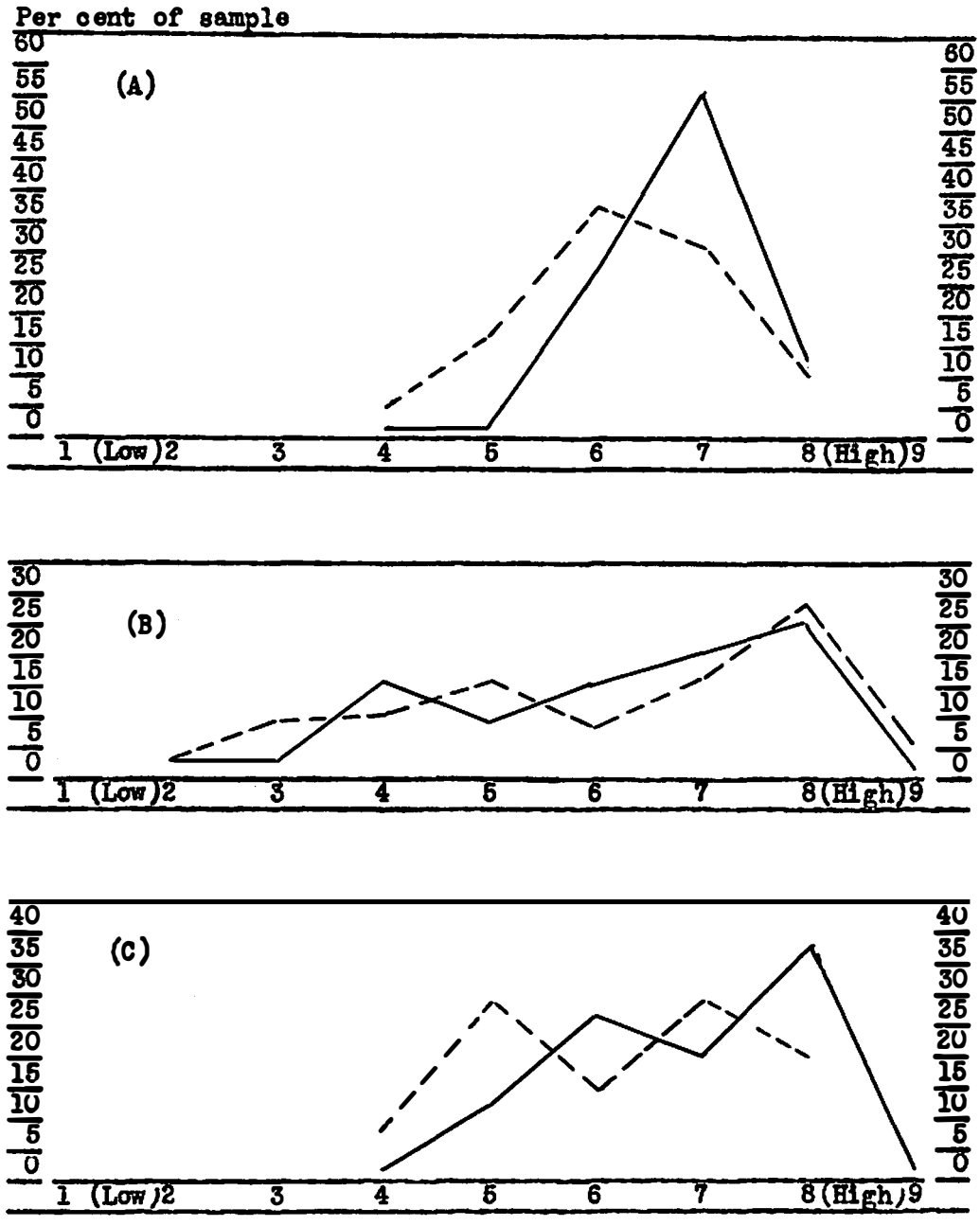


Figure 12. Comparison of criterion scale results by sample used, person and effectiveness criterion scales; _____ = person, ------ = effectiveness.

A. Validation Scale Distribution, using class average rating. (N = 71.)

B. Qualification Check List Distribution, individual student ratings. (N = 3600.)

C. Validation Scale Distribution, individual student ratings. (Validation sample of 100.)

by per cent of sample for both criterion scales in the QCL administration (B in the figure); and the distribution for both criterion scales for a sample of 100 Validation Scales used for calculation of item correlations in the cross validation sample (C in the figure). A random sample of 100 scales rather than all 1900 scales was used to avoid weighting by school or department, and to reduce the data to a manageable size (using calculating machine rather than IBM methods). The 71 classes were arranged in the order the results were received; then, using a table of random numbers, one scale was selected from each class in turn. When 71 such scales were thus collected, additional scales were added through use of the table until 100 scales were obtained. This sampling procedure undoubtedly created some artificial restriction in the distributions shown in part C of Figure 12.

The difference between the QCL and Validation Scale results is seen by contrasting A and C (Validation Scale) with B (QCL) in Figure 12. The Validation Scale results are markedly restricted for both person and effectiveness criteria, but the distributions are more nearly normal. In considering the effect of these criterion scales on item validity indexes, it should be kept in mind that the QCL correlations were based on a 5 by 9 point bivariate distribution, whereas the Validation Scale had a 15 by 9 point structure. In spite of this, when the item effectiveness correlations (Z') obtained from the QCL were plotted against the discrimination indices obtained from the Validation Scale, a linear relationship and a product-moment r of .78 was obtained. From this it would appear that, generally, the item-scale relationships are relatively

stable from one administration to the other, since the obtained correlation is undoubtedly attenuated. The relative magnitude of validity indexes in the two administrations is another question, to be reported later in this chapter.

Figure 12 has also been constructed to show how students handle the requirement of rating on both liking as a person and teaching effectiveness. In spite of the moderately high correlations obtained (.73 in the pilot study, QCL: .54 in the QCL administration; .64 in the Validation Scale administration, using mean ratings,) between person and effectiveness ratings, it can be seen that in all cases (A, B, and C) there is a meaningful difference between the two criteria. It is interesting that this difference is more marked in rating "real" teachers than in rating "a" teacher, as in the QCL. This is seen by inspecting A and C differences in person and effectiveness distributions in contrast to B.

Contrasting "effectiveness" in A and C, it would appear that, although the surfaces are not parallel, there is a rough equivalence; enough so that the sample of 100 can be accepted as crudely representative of all 1900 ratings for the purpose of computing item indexes.

Summary of Validation Statistics

Because at least a rough stability and equivalence of criterion data has been established, a comparison between item indexes obtained

from the cross-validation sample and the QCL sample may be made. The discrimination and biasability indexes (r_{ie} and r_{ip}, respectively,) from the QCL and Scale administrations for all one hundred and sixteen items are presented in Appendix J. The "key" or higher discriminating item is marked by a parenthesis and the couplets are separated by a space. All entries are in Z'.

From this table, Figure 13 has been prepared to show the effect on discrimination indexes (r_{ie}) of the administration of the scale to a cross-validation sample. It will be recalled that a 15-point scale is being used on the item (instead of a 5-point as in the QCL) and that the criterion variance is restricted. Therefore, the observed difference in distribution of the indexes cannot be attributed solely to administration of the scale in a practical setting, i.e. using a named teacher. Whatever the reason, it can be observed that the scale indices are substantially lower. The mean difference is about 25 Z' points. Only one of the 116 items had a higher discrimination index in the scale administration.

Figure 14 shows the results for the biasability index, r_{ip}. In this case, eight of the one hundred and sixteen items had higher Scale indexes. The algebraic mean difference is 18 Z' points.

However, a drop in <u>each</u> item discrimination or biasability index in the Scale administration does not necessarily mean a corresponding drop in couplet or item-difference score. For example:

Item	QCL	Scale	QCL-Scale Difference
3	89	76	13
4	63	40	23
Couplet Score	26	36	10

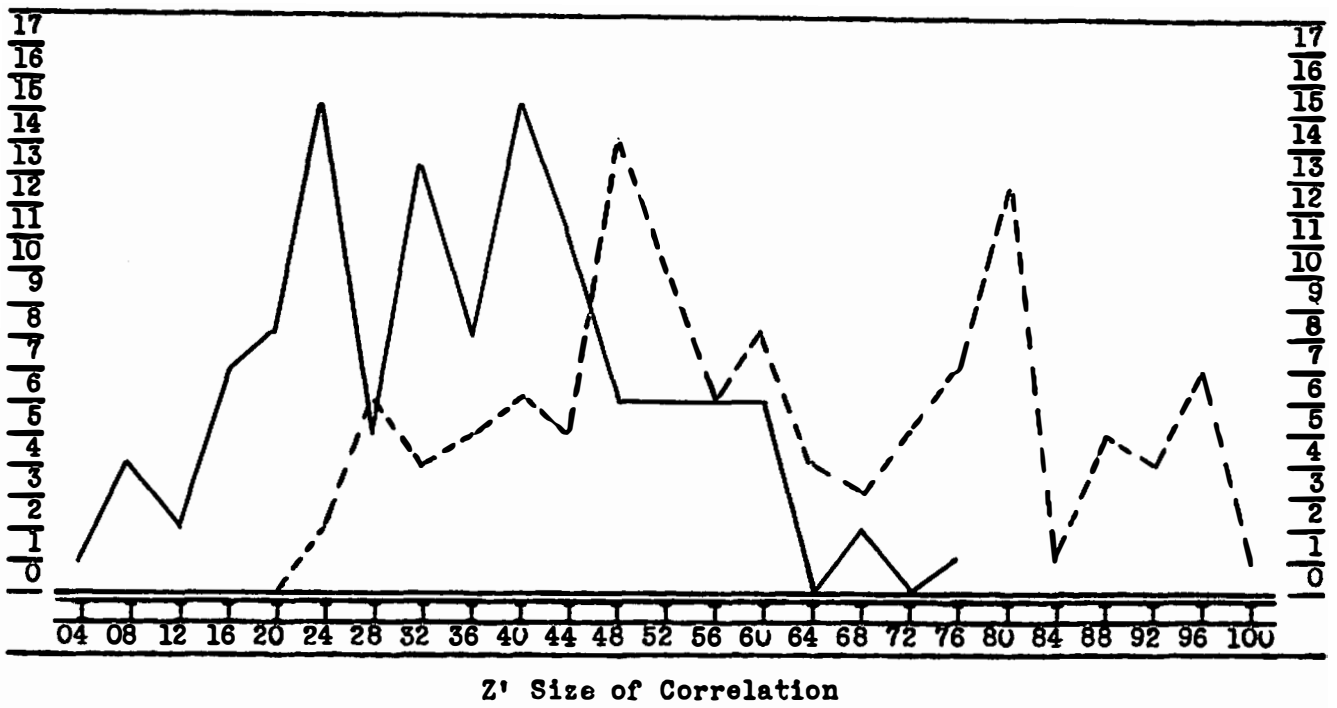


Figure 13. Distributions of discrimination indexes (r_{ie}) for Qualification Check List and Validation Scale. _____ = Validation Scale. ------ = Qualification Check List.

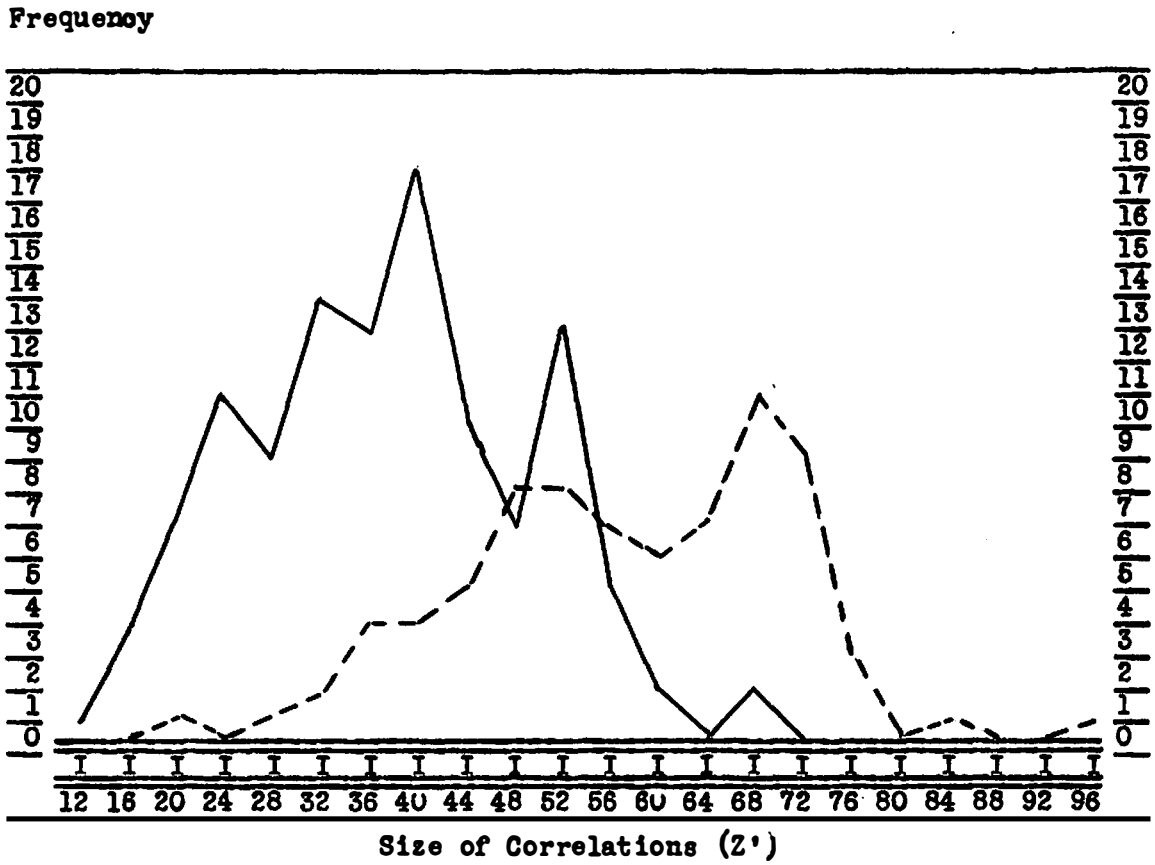


Figure 14. Distributions of biasability indexes (r_{ip}) for Qualification Check List and Validation Scale. ——— = Validation Scale. ------ = Qualification Check List.

Here, although item 3 dropped 13 points and item 4 dropped 23 points, the difference or couplet score increased 10 points in the Scale over the QCL. But this is not characteristic. Figure 15 shows the distribution of item difference scores within couplets for the QCL and the Scale. The average difference score for the QCL was 32; for the Scale it was 21. There were only 8 cases, as in the example above, where Scale difference score was higher. The algebraic couplet mean QCL-Scale difference score was 14; that is, on the average the QCL difference score between items was 14 points higher than in the Scale. Since we are dealing with discrimination indexes, this is disturbing. This kind of loss is significant and does not appear to be readily explained in terms of regression effects. It seems that this is "real" loss, attributable to using a named teacher. Figure 15 also shows another kind of loss, one in which the item higher in discrimination in the couplet in the QCL does not hold up as highest in the Scale administration. These are the minus Z' values in the figure.

The biasability index difference within a couplet should, of course, be as low as possible, with zero difference as the objective. Figure 16 shows the distribution of these differences for both QCL and Scale administration. If not zero, then the direction of the difference should be as shown in the following example:

(key item)	Item x	High Discrimination	Low Biasability
	Item y	Low Discrimination	High Biasability

This follows because choice of item x as most characteristic by the rater can then be said to be against the pull of a high bias factor; if y is chosen the rater is saying the nice thing. Obviously, this has

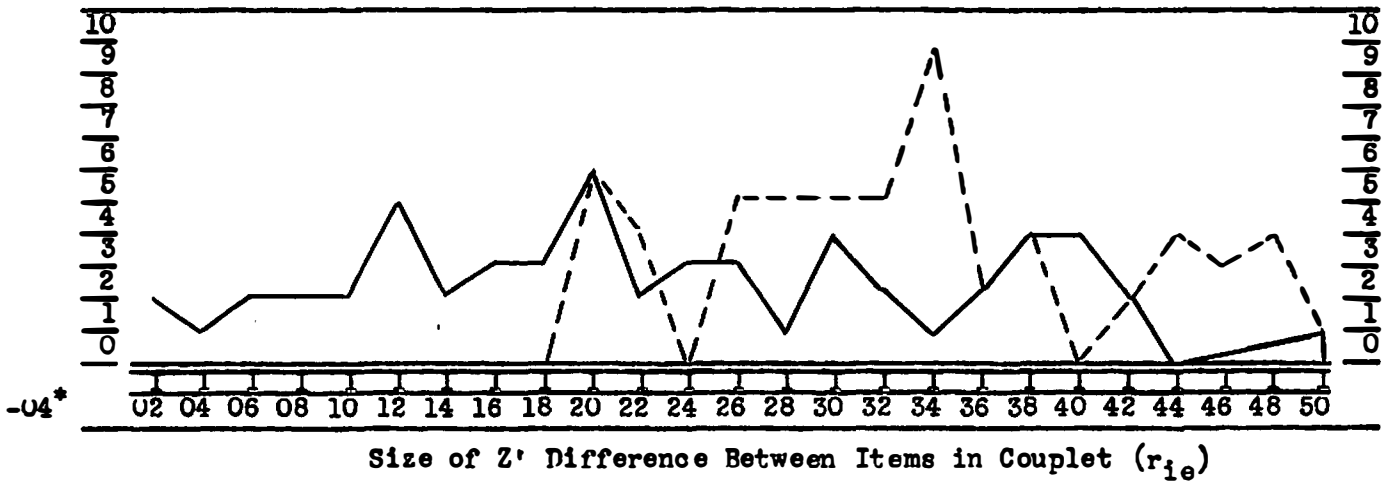


Figure 15. Distributions of discrimination index difference within couplets for items on Qualification Check List and for Validation Scale. _____ = Validation Scale. ------ = Qualification Check List.

* Three couplets in Scale administration had differences in the opposite direction to that obtained in the Qualification Check List. Size of difference: -04; -06; and -16.

limits, and extremes are to be avoided. The case cited is defined as negative in Figure 16. The case where the discriminating or key item in the couplet is also higher in bias or correlation with person rating is defined as positive.

It will be observed that there are more minus than plus couplets, and that the Scale contains more positive and more negative couplets than the QCL. The mean minus value for 33 QCL couplets is 7.8; for the Scale, 42 with a mean value of 12.2. The mean positive value for 20 QCL couplets is 4.4; for the Scale, 13 positive couplets have a mean of 9.5.

In all of the above discussion, it should be recalled that a sample of 100 scales was used to calculate scale indices. Results obtained from all 1900 cases could be quite different.

Couplet Validation

The point in identifying the items which are unstable is that of eliminating them from the final scale or at least eliminating them from the scoring of the final scale. However, this procedure by itself is not sufficient. The difference in the mean on the 15-point scale between items in the couplet (couplet mean difference) was correlated against the effectiveness criterion scale score and used as an estimate of <u>couplet</u> validity. Validities for each of the fifty-eight couplets will be presented in Chapter IX, Construction of the Forced-Choice Scale. Table V summarizes overall results by Pearson r, not Z'. This table is interpreted by the finding that the couplet validity correlation is significant

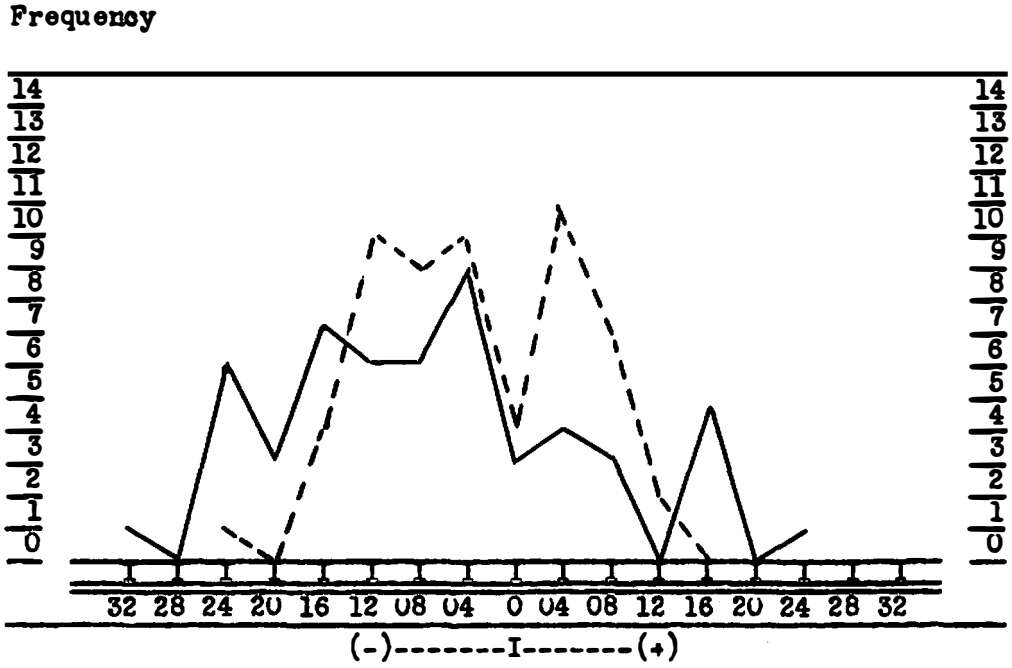


Figure 16. Distributions of biasability index difference within couplets for Qualification Check List and Validation Scale. _____ = Validation Scale. ------ = Qualification Check List.

at .05 probability (one sided alternative) if the r is equal to or greater than .197. On this basis 11 of the 58 couplets, or about 19 per cent, were not usable.

Departmental Differences

Couplet validities were obtained across all 71 teachers and 9 departments. The question arises whether or not these couplet validities would hold up in every department. In order to investigate this problem, it is necessary that each department be a sufficiently large and unbiased sample of the total group with respect to the criterion, the effectiveness scale. Tables VI and VII show the results of an analysis of variance of the mean effectiveness score by department. Because of the widely different number of teachers in each department and the effect of this on normality and trait variance, the F ratio rejection at the 5 per cent level of the hypothesis that the departments are drawn from a common population is suspect as a generalization. The Hartley F maximum test shows the variance as significantly heterogeneous and provides another reason for questioning a probability of .05.

Differential Scoring Analysis

It will be recalled that the Validation Scale was administered as a "disguised" forced-choice scale, that is, couplets were built into the

TABLE V

DISTRIBUTION OF COUPLET VALIDITY CORRELATIONS BY SIZE

Size of Correlation (r)	Frequency
Negative	2
0 - 19	9
*20 - 29	15
30 - 39	13
40 - 49	10
50 - 59	8
60 - 69	1
Total	58

* All correlations this large or larger significant at probability .05

TABLE VI

ANALYSIS OF VARIANCE, MEANS OF EFFECTIVENESS SCALE BY DEPARTMENT

No. of Classes (Teachers)	Department Code Numbers								
	A	B	C	D	E	F	G	H	I
1	6.16	5.39	4.85	6.34	6.06	5.84	3.97	5.79	6.79
2	6.09	5.86	6.57	4.40	5.74	3.50	5.94	4.51	5.71
3	4.44	4.68	5.40	5.00	5.29	3.09	3.95	5.31	6.44
4	5.83	4.81	6.67	6.69	5.00	3.55		4.53	5.78
5	4.20		5.39	4.11	5.39			5.10	
6	5.09		6.60		4.07			6.00	
7	5.15		3.46		4.64			4.57	
8			6.41		5.57			6.14	
9			6.50		5.10			5.41	
10			6.90		4.52			4.75	
11			5.00		7.43			5.96	
12			6.62		4.84			4.62	
13					4.27				
14					6.00				
15					4.39				
16					5.89				
17					6.28				
18					4.59				
19					5.17				
20					5.69				
Sum	36.96	20.74	70.37	26.54	105.9	15.98	13.86	62.69	24.72
$\bar{X}$	5.28	5.18	5.86	5.31	5.30	3.99	4.62	5.22	6.18
S	.43	.28	1.05	1.04	.59	1.21	.88	.40	.21

TABLE VII

VARIANCE TABLE

Source Variation	Sum Squares	df	Mean Square	F
Between Departments	15	8	1.8	2.368
Within Departments	47	62	.76	
Total	62	70		

scale but were not revealed to the student. The assumption being made here is that a student would, if forced to choose as in forced-choice format, select that item of each pair to which he would give the higher scale value. As established in the section on couplet validation, the couplet mean difference score is valid for most of the couplets.

Going ahead with the assumption that a forced-choice score is essentially a difference score permits an analysis of the effectiveness of various scoring methods in predicting the two criterion scales or a composite of them. Table VIII reports validities of the Scale obtained by various scoring methods.

A score for each teacher by method labeled 1a in Table VIII (Forced-Choice-algebraic, unweighted,) was obtained by counting the number of times the highly discriminating item (the "key" item) in the couplet had a higher value. If higher, a positive value was assigned. This is all that would be known if the directions to the student had read, "choose the most characteristic of the two items;" this is the "as-if-forced-choice" method. Scores thus derived for the 71 teachers were correlated against the effectiveness criterion scores and the person criterion scores. A scatter diagram showed a linear relationship. In the procedure employed, the difference score was obtained algebraically and then a constant was added to remove negative values.

In method 1b, weighting by Z' difference, a scale for each teacher was overlaid with a key showing the Z' difference between the key and non-key items in the couplet. When the key item mean score was high, the Z' weight was positive; when low, the Z' weight was negative. The

TABLE VIII

VALIDITIES OF SCALE SCORED BY DIFFERENT METHODS

Scoring Method	r. Effectiveness	r. Person
1. Forced-choice		
a. Algebraic, unweighted	.45	.66
b. Weighted by Z' Difference	.52	.18
2. Graphic, Non-Key Items	.78	.77
3. Graphic, Key Items	.91	.92
4. Graphic, All Items	.91	.84
	Composite Criterion, E>P = +	
5. Forced-choice		
a. Unweighted	r = .77	
b. Weighted by Z' Difference	r = .70	
6. Graphic, All Items	r = .53	

score for the teacher was then derived by adding the Z' values for + couplets and the Z' values for the - couplets. A constant was then added to convert to positive values.

In method 2, the mean score for each item in the scale which was not a key item (the key item was at least 20 Z' points higher than its opposite number in the couplet) was added and divided by 58 (the number of non-key items) so as to obtain a mean score for each teacher. In effect, this cuts the scale in half, using the lower discriminating items. In method 3, this was reversed, using the key items.

In method 4, the grand mean of all 116 items was used for each teacher. In methods 5 and 6, the criterion score was changed to a composite one, so that a difference score between the mean effectiveness and mean person scales was considered positive if the effectiveness mean was higher. The thought here is that the real concern is one with teaching effectiveness beyond that to be expected because the teacher is liked. (It will be recalled that roughly 50 per cent of the variance in effectiveness is not predicted by person rating.) The weighted and unweighted methods in the forced-choice scoring were the same as with the single rather than composite criterion.

Various other methods of scoring were considered and postponed for future research. In the next chapter, in discussion of the forced-choice scale, the statistics presented here, plus some related ones, will be brought into perspective.

CHAPTER IX

CONSTRUCTION OF THE FORCED-CHOICE SCALE

Introduction

In order to explain effectively the procedures used in constructing the final form of the scale, it is necessary to consider its future research use, which is a matter more properly presented in Chapter X. Briefly, then, it appears that as a result of this study a rather definitive method of appraising the effectiveness of forced-choice versus graphic rating scales for student evaluation of college teaching can be formulated. It is proposed that two scales be administered to the same sample of teachers, half of each class using a graphic scale to rate the teacher and half of the class using a forced-choice scale. The criterion would be the same two criterion scales used in this research, as well as any "external" criterion obtainable.

In such an experiment the necessity of equal treatment of all variables in the scales except forced-choice format is clear. The scales should be of equal length, content, and structure. For this reason, and in order that the methods reported in this thesis may be replicated, the forced-choice scale developed here consists of the complete Validation Scale, altered to conform to forced-choice format. However, not all the couplets in the forced-choice scale so constructed need be scored. The next section develops specifications for scoring.

Development of the Forced-Choice Scoring Key

It will be recalled that three indexes and two administrations were used in cross validation of items. These variables are shown in the heading of Table IX, Summary of Cross Validation Results.

The first step in selecting couplets for scoring in the final scale was to isolate those couplets which had a difference in discrimination index (r_{ie}) between items, in both QCL and Scale administrations, equal to or greater than 20 Z' points. Applying this selection procedure resulted in retention of 32 of the original 58 couplets. (Table IX, Column 2.)

Next, using the biasability index (r_{ip}), those couplets were selected which had item differences in both administrations of <u>less</u> than 10 Z' points. Twenty-seven couplets were so obtained. (Table IX, Column 3.)

When both of the above criteria were employed simultaneously, (a most rigorous selection procedure,) only 14 of the 58 couplets remained.

The mean of the items in these couplets in the Scale administration were then inspected to determine if the characteristicness indices were approximately equal. In these 14 couplets in all but two cases the non-key item (one with lower r_{ie}) had a higher mean. Difference magnitudes varied from 0.3 to 1.00 scale point, whereas .6 of a score point (on the average) was required at the 1 per cent level for <u>one</u> item in rejecting the null hypothesis, so that these differences in means are not likely to be significant. Table IX shows these mean differences

TABLE IX

SUMMARY OF CROSS VALIDATION RESULTS

	Couplet Number	Item Numbers	r_v	(1)	(2) Z' Difference Discrimination QCL	(2) Z' Difference Discrimination Scale	(3) Biasability QCL	(3) Biasability Scale	(4) $r_{ip}-r_{ie}$ Difference QCL	(5) $r_{ip}-r_{ie}$ Difference Scale	(6) $D_{\bar{x}}$ QCL	(7) $D_{\bar{x}}$ Scale
1.	8	(15)16	.36		.27	.21	-.06	0	.21	.21	.88	-.023
2.	2	(3) 4	.52		.26	.36	.07	-.07	.33	.29	.84	.013
3.	12	(23)24	.35		.33	.29	.05	.03	.38	.32	.68	.034
4.	11	21(22)	.45		.47	.38	-.02	-.01	.45	.37	.88	-.025
*5.	13	25(26)	.36		.50	.31	-.07	-.29	.43	.02	.90	-.027
6.	14	(27)28	.34		.44	.24	.01	-.07	.45	.17	-.13	0
7.	17	33(34)	.46		.45	.39	-.12	-.17	.33	.22	1.11	-.126
8.	18	(35)36	.52		.30	.29	-.07	-.16	.21	.13	.40	-.039
9.	20	(39)40	.35		.32	.25	-.09	-.04	.23	.21	1.03	-.028
10.	22	(43)44	.53		.38	.41	.02	-.26	.40	.25	.81	.003
11.	24	(47)48	.24		.46	.35	.08	-.09	.54	.26	.59	.003
12.	25	49(50)	.41		.33	.23	-.10	-.14	.23	.09	.62	-.017
13.	26	(51)52	.23		.27	.27	.10	-.03	.37	.24	.86	0
14.	27	(53)54	.25		.36	.20	.10	-.22	.46	-.02	.61	.255
15.	29	57(58)	.30		.31	.37	-.07	-.12	.24	.25	.48	-.008
16.	30	(59)60	.45		.20	.20	.03	-.06	.23	.14	.43	-.033
17.	33	65(66)	.20		.20	.20	-.06	-.07	.14	.13	.77	-.049
18.	37	73(74)	.27		.29	.20	.02	-.03	.31	.17	.93	.028
19.	38	(75)76	.34		.34	.29	-.10	-.06	.24	.23	.35	-.018
20.	41	81(82)	.53		.48	.40	-.09	-.16	.39	.24	.54	-.009
21.	42	(83)84	.27		.41	.21	-.13	-.24	.28	-.03	.13	.008
22.	47	93(94)	.27		.31	.33	-.09	-.12	.22	.21	.32	-.007
23.	48	(95)96	.56		.35	.50	.04	-.17	.39	.33	.76	.045
24.	50	(99)100	.42		.46	.38	0	-.24	.46	.14	1.08	.020
25.	51	101(102)	.22		.38	.32	-.12	-.08	.26	.24	.25	0
26.	52	(103)104	.33		.25	.41	-.15	-.23	.10	.18	-.17	-.011
27.	53	105(106)	.28		.44	.24	.07	.16	.51	.40	.77	-.021
28.	55	109(110)	.33		.29	.30	-.09	-.12	.20	.18	.61	-.010

r_v = Validation correlation (mean difference against Effectiveness Scale).
( ) = Key item.
- signifies key item with higher r_{ip} than non-key.
$r_{ie}-r_{ip}$ Difference is algebraic.
* Example discussed in text.

for the QCL as well as the Scale administration. By inspection it was concluded that the QCL differences were probably not significant.

Up to this point couplet validities (as opposed to item difference consistencies, had not been considered, although they had been calculated as described in Chapter VIII. Accordingly, the 47 couplets with validities of .20 or greater were inspected for consistent discrimination index of 20 Z' points or greater and those meeting this criterion were included in Table IX. Thus 28 couplets are shown in Table IX. The couplet number in this table was taken from the worksheets; item numbers are shown on the printed scale, and the "key" item is placed within parenthesis.

In this last procedure the biasability index had been ignored, so the size and direction of the difference was entered in Column 3. Entry 5 in the Table, couplet No. 13, is taken as an example. This couplet appears in the scale as:

25. Gives regular quizzes.
26. Inspires you to make the maximum preparation for each day's assignment in class.

In the QCL, Item 26 was 50 Z' points greater in discrimination than Item 25; in the Scale it was 31 Z' points greater. However, in the QCL, Item 26 was also greater in correlation with the person scale by 7 Z' points than Item 25, i.e. biasability in the same direction as the more discriminating item. Since this is considered undesirable, a negative sign is given this seven point difference. In the Scale administration the situation is worse; at -29 points, Item 26 differs from Item 25 almost as much in biasability as discrimination and in the same

direction. If a student checked Item 26 in this couplet as "most characteristic," he may have done so as much on a liking basis as on considerations of effectiveness.

Because the topic is one of developing a scoring key, rather than exclusion of the couplet from the scale as administered in the future, this situation is not regarded as critical at this point in the program of research. Various scoring keys can be used in the forced-choice scale and their effectiveness established. In this connection it should be kept in mind that a random sample of only 100 cases from the 1900 scales obtained was used to calculate these Scale indexes by the method described in Chapter VIII, and that differences are being discussed. It is regretted that IBM equipment was not available so that the whole sample could be used. Therefore, considerable fluctuation can be expected. Taking the obtained figures at face value, it would appear that 10 of the 28 couplets listed in Table IX should be eliminated from the scoring key. One of the advantages of the experiment proposed in the introductory section would be the opportunity it affords to validate couplets in forced-choice format and to cross-validate couplets in the graphic scale.

In summary then, it is suggested that, in the interest of reliability, all 28 couplets be scored in the forced-choice version of the scale and that the data be checked as outlined in the preceding paragraphs. If IBM equipment is available, it is further suggested that Brogden's formulas for distortion score and distortion index be used as part of the experiment.

The Forced-Choice Scale

The content of the scale is presented in this section, not the scale itself. Because the scale should be printed on the same size paper and present much the same appearance as the Validation Scale, there appeared to be little point to working out the mechanics of an adaptation as part of this thesis.

Starting with the directions to the student, then, the Forced-Choice Scale will take this form:

Directions

> Please describe the teacher by selecting one statement in each of the pairs of statements below. Indicate which of these statements is most characteristic or typical of this teacher. Be sure to read both statements carefully. Even though both statements apply, select the statement which most often or most strongly applies. Describe the teacher in this way as accurately and honestly as possible. Please do not omit -- guess when you must. Do not let anyone else influence your judgment and do not sign your name. Put an "X" in the appropriate box under Choice, below.

The body of the scale will have the following format:

	STATEMENTS	CHOICE
1A	Good posture and bearing.	
1B	Uses sufficient supporting details in lesson.	

The criterion scales will appear in their current format following couplet Number 58 in the body of the scale.

The directions in the scale are quite uninspired. It would be rewarding perhaps to experiment with directions, but the restriction

to "standard" forced-choice format is necessary if comparisons are to be drawn.

CHAPTER X

DISCUSSION OF RESULTS AND PROPOSALS FOR FURTHER RESEARCH

Introduction

It may have been noted that, in the presentation to this point, certain topics conventional in a discussion of rating scales have been omitted. The considerations involved in student rating of college teachers were discussed in Chapter II, and the rationale of forced-choice scale construction was presented in Chapter III, but topics common to all rating scales, such as validity, reliability, bias, and leniency, have been postponed to this chapter in order that they may serve as a framework for the discussion.

This chapter, then, is organized around objectives or goals of rating scale construction and use. The results of this research in attaining these objectives and proposed steps in a program of further research will be discussed.

Validity

It should be understood at the beginning of this discussion that only two of the possible relevant criteria have been used in this research; those of student opinion of the teacher's effectiveness and of his status as a person, as defined by the two nine-point criterion scales used. Some students of the problem of student rating of teachers are not

satisfied with such criteria. They wish something beyond student judgment and ask the "really" question, "Is student opinion really a good criterion; how do we know that the student definition of good teaching really defines good teaching?" This school of thought requires an "external" criterion; for example, "student growth" (although the more practical say student grades). The difficulty with such external criteria is this: they are relevant not only to teaching but to a host of other variables as well, so that it is quite impractical to design an experiment and generalize from data collected with regard to the effectiveness of teaching.

However, a sort of bootstrap lifting can be done. The scale score could, for example, be used as a criterion for a variety of instruments in selecting teachers. Suppose it could be established that interest in teaching (as measured, say, by the Strong Vocational Interest Blank), intelligence (as defined by the Wechsler-Bellevue), scholarship in the field taught, and a number of other variables, all correlate significantly and fairly well with student opinion as revealed by the scale. In the absence of negative information it would seem reasonable to conclude in this case that the effectiveness scale criterion is satisfactory. The person scale would be retained and results compared with the effectiveness scale in each such case. It is proposed that the functional utility of the scale be checked in this fashion in further research. If "external" criteria can be obtained which are unambiguous and directly related to teaching effectiveness, they should of course be used also.

In refining and developing criterion data, the question of assigning weights to reduce criterion distortion should be resolved by study of the article by Ryans (26). Apparently arbitrary weighting such as is used here is as good as any.

Turning to the validities of the scale as reported in Chapter VIII, Table IV, the fact that the Personnel Research Section, Department of the Army, obtained "equally good validity" by using graphic rather than forced-choice rating may be recalled. In this research, graphic scoring, considering only the effectiveness scale, produces appreciably higher validities, the best forced-choice method correlating .52 and the best graphic method .91. It should be noted, however, that the graphic score also correlates very highly with the person score, whereas the forced-choice score (particularly the score derived by weighting the Z' difference) has a lower correlation with the person scale. This is seen in another way when the composite criterion is used. A high correlation here indicates high net effectiveness; effectiveness over and beyond that to be expected on the basis of liking for the teacher as a person. On this basis the forced-choice method of scoring appears to be the better predictor.

It is quite apparent, however, that much yet needs to be done to clarify the relationships between items, couplets, and criterion scales. It is proposed in this connection that the class mean item difference score itself be used in scoring, both weighted and unweighted by Z' difference, to see if this forced-choice scoring has an even greater differential validity, i.e., is a better predictor of effectiveness than

liking. Before more work is done in this area, however, the meaningfulness of the effectiveness-person difference as a criterion has to be established on other than a priori grounds. Does the teacher who is rated higher on the effectiveness scale than on the person scale show up as consistently higher on the effectiveness scale itself? A scatterdiagram might prove informative here. Moreover, a product rather than a difference composite criterion may be more useful both in maximizing individual differences and on theoretical grounds. The concept here is that a teacher is successful to the extent he is able to influence others and is more likely to succeed if he is liked, provided that he is also competent as a teacher. Another scoring method of potential promise is deviation scoring, using item mean and sigmas and criterion mean and sigmas. Yet another method might be a weighted effectiveness score in a composite criterion and subtraction of the person score; thus: (2E-P).

In the discussion so far, the implicit assumption has been that it is necessary to find the best predictor of the criterion scale or scales by developing the most efficient scoring method. Several things are wrong with this. First, there appears to be no pressing reason to eliminate the criterion scales in operational use as part of the scale. If the criterion itself is available, then why the need for a predictor? Second, emphasis is then placed (retaining the assumption) on an overall evaluation or summary figure rather than on informing the instructor of specific areas in which students say he needs to improve. A next step in the research might well be a factor analysis of the one hundred and sixteen item scale; or, in the interim, a part-scoring of the scale,

using the categories originally set up in the Qualification Check List. In this connection a particularly valuable appraisal of the effect of forced-choice methodology may be made. Since part scores should have sufficient range to be reliable, an essential step here is that of taking the original category distribution of the three hundred and ninety-four items and seeing how many items remain in each category. Next, the one-hundred-and-sixteen item scale should be analyzed in similar fashion. The validity of the scale may rest too heavily on one phase of classroom work and there may be too few items in a given category,

Validity has many connotations, but central in the concept is the notion of utility for a specific purpose. This utility is expressed, if possible, in quantitative terms, but the crucial aspect is that of the objective. The best possible evidence of the validity of the scale produced in this research would be that of establishing its utility in improving instruction. Therefore, it is proposed that longitudinal studies be made so that instructor growth may be estimated. The administration of scales to a control group of teachers who do not see the results is only one of a variety of problems here; student and teacher "adaptation" would need to be controlled, and the problem of a criterion external to the scale itself needs to be solved eventually.

Somewhat peripheral to this kind of validity study but related to it is an appraisal of the <u>need</u> for such a scale. Instructors should be asked to predict class response to the scale. These predictions can then be compared with class results, item by item. If high correspondence exists, one possible conclusion is that little need exists but that the

scale is highly valid, since the teacher and class agree! On the other hand, if little agreement is found, the way is then open to an "action research" approach to improvement, both of the evaluation instrument and of the evaluation method. A classroom critique and analysis, item by item, of the results of class and teacher administration (teacher not present) and then a later interview (or series of them) with the instructor on the part of the researcher, would not only serve to establish the kinds of instructor (and class) needs, but would also provide a valuable method of item refinement for future scales. Naturally, individual differences in the ability of instructors to predict class response would be analyzed. A substantial correlation between such an ability and student ratings would afford further evidence of scale validity. As a variation on this theme, teacher self-ratings may be used, using a different group of teachers.

Another aspect of validity may be mentioned -- that of item validities under cross validation. Langmuir (13) in dealing with cross validation states that, "Empirical studies and particularly cross validation (studies) have an insidious way of destroying confidence in systems of evaluation and prediction." Loss of items, then, is expected, but the loss here of all but 14 out of 58 couplets because of item-criterion correlation changes between the QCL and Scale administrations appears to be excessive in view of the use of some 500+ students for each form of the QCL. One probable reason for the loss is the use of a sample of 100 students out of a total of some 1900 to compute the item correlations. It is proposed that, when IBM equipment can be obtained, the item

correlations be run using the entire sample. Also, it is suggested that, instead of using individual student ratings, the mean score by class of each item be correlated against effectiveness and person mean ratings over the 71 teachers in the sample. This last suggestion follows from the fact that, since shrinkage is expected because of operational rather than research administration of the scale, it is therefore logical to use the operationally significant value, the item mean. The research proposed in Chapter IX, Construction of the Scale, with respect to parallel administrations of the two forms, forced-choice versus graphic, would of course provide further information on item and couplet stabilities, as well as provide a test of the validities of couplets in forced-choice format.

Suppressor Variable Proposals

In planning further research to investigate and clarify the relationships among items and criterion scales, particularly with respect to using mean ratings, it is recommended that Brogden's approach be integrated into the design of the forced-choice versus graphic format experiment. It may be recalled that the use of average ratings, according to Brogden (3), results in loss of individual differences in distortion, and invalid variance is not reduced. In the discussion to follow, it should be noted that it is not necessary to collect additional information, as in the experiment proposed above, in order to apply the suppressor variable approach to the scale built as a product of this thesis research.

The rationale is based on the general concept as presented by

Loevinger (16) and by Brogden (3) as previously cited. To the equation $X_i = T_i + E_i$ (where X_i is the score of i individual on test X, T_i is his "true score" and E_i is the random error) Loevinger adds D, signifying error correlated with the true score. D represents distortion of measurement. Formulas are presented by Loevinger to the effect that the test (scale)-criterion correlation is reduced "far more sharply" by distorting factors than by random error factors. Loevinger also presents Brogden's (3) distortion index formula in her paper, using her symbolism.

The procedure, following Brogden, is to obtain a couplet by pairing items with the largest possible difference in discrimination (r_{ie}). However, instead of the r_{ip} used here, Brogden suggests using equal item-distortion score correlations as a control device. Since these correlations are equal, the response to the couplet should be uncorrelated with the distortion score, i.e., should be unbiased. The reasoning is that the response to the couplet is essentially a difference score, hence it is not correlated with the distortion score because neither of the items in the couplet are so correlated.

The problem, then, is to obtain a distortion score against which to correlate each item. In his paper, Brogden (3) uses an "early version" which turned out to be impractical and theoretically difficult to make the notion of distortion clear. In this procedure, a scale would be administered twice to the same sample of students, once under "actual selection" conditions and once under "frank" conditions. The difference in the scores (ratings) an individual assigns in the two administrations is the distortion score for the item in question. Among the defects of

this method, Brogden (3) cites the memory factor, the difficulty of frank responses being frank, and the limitation to bias of conscious origin.

As a result, Brogden proposes that the score X, on an item, i, be regarded as composed of the sum of the following variances: y, a component common to key and criterion; d, a distortion component; s, components specific to the item other than distortion; e, error. The variances s and d are to be treated as invalid. If, then, y can be estimated, it is possible to treat the distortion score as $d_i + s_i + e_i$; and to obtain it by substracting an estimate of the score component, y, from x_i. To estimate y, the regression of the test on the criterion is used to predict the test score, and disregarding the constant, the distortion score is estimated from the formula, $x_i - b_{xo}c_i$. The distortion index is obtained by a biserial correlation between each item and the distortion scores for the individuals. Biserial correlation is used because Brogden is talking about questionnaires in "yes-no" form; as applied here product-moment correlations would be used. Two samples, one for item validities and one for distortion index purposes under operating conditions, would be required.

It is recommended, therefore, that the data gathered in this research be treated as outlined above and the results compared with the data of this thesis. The QCL sample can be split into two subsamples, one of which can be used to compute the distortion index and the other to recompute the validity indices. The necessary statistics can be obtained from the data collected. It is probably unnecessary to add that

IBM equipment would be required to compute approximately 1900 distortion scores!

If the forced-choice and graphic versions of the scale were administered as proposed previously, a test of the effect of forced-choice format on the correlational properties of the items and pairs could be made. Also, and following Brogden's recommendation, the graphic scale, if both scales were built on suppressor variable principles, could be scored so that a set of suppressor items with high distortion indices but with approximately zero validities is assigned negative weights. Or, alternatively, the scale could be scored using only those items not subject to distortion. As Brogden points out in this connection, "Many subsidiary hypotheses and assumptions in the rationale and procedures have not been directly tested and proved." It appears that the work reported here affords an opportunity to make some of the relevant tests. Particularly interesting to the writer would be an inspection of the relationship between the distortion index and the biasability index of this thesis.

Reliability

No reliability studies have been made in this thesis. Quite obviously, before the scale is to be placed in operational use such studies must be made. In view of the overriding importance of validity as opposed to reliability, and the financial and time costs involved, it was decided to place reliability in second place. The next step in the work projected for the future as opportunities become available is

that of studying item and scale reliabilities.

Considerations in choice of a method for estimating the reliability of a rating scale are somewhat complex. On the one hand, we are interested in the behavior of the item as it reflects class opinion at any given point in time. For this purpose a measure of dispersion or scatter on the item among raters, such as the standard deviation, is probably as significant as any other statistic. From another point of view, the stability of the item over a period of time comes into question. This involves using the same raters and same scale, making the assumption that neither growth of teacher nor student memory are material factors and obtaining test-retest correlations.

Turning from items to the entire scale, the intraclass correlation, or average agreement of all raters over all items, is probably appropriate. The various odd even or Kuder-Richardson formulas do not appear to be especially meaningful or appropriate.

Student, Teacher, and Instructional Variables

In order to shorten this discussion as much as possible, the reader is referred to the section of the QCL headed Classification Data (Appendix C) for a description of data available in investigating the effect of student, teacher, and teaching situation on the results. Theoretically, an analysis of variance for each individual item should be made for these variables. Practically, a comparison of item means and standard deviations for various groups and situations is a more feasible approach. The objective of course is to find out which items

are valid for all groups and then to decide what unique sets of items can be used as supplementary scales. The final product would be a master scale composed of universal items and a series of supplementary scales to be used in specific situations as required. Basic to such an undertaking, of course, is some method of deciding how much predictive or diagnostic scale utility is sacrificed by not using supplementary scales. An analysis of variance would be useful in this connection. It should be noted that a factor analysis of the item intercorrelations is something different and probably should be done after the set of "universal" items has been isolated.

Related Technical Issues

The effectiveness of couplets constructed using the characteristicness index together with the biasability index should be compared with their effectiveness using the biasability index (r_{ip}) alone. The method employed by Berkshire (2) and reported in Chapter III appears adequate for this purpose. It may be that use of the characteristicness index is inefficient. Its elimination would be one less restraint and more couplets could then be built from the QCL data.

The question of bias, and especially leniency, is amenable to attack through use of information collected through the use of the administration log in the scale administration. Part of the cross validation sample was administered the scale with instructions to the student that the purpose of rating was to help the teacher improve; part of the sample received this instruction but was also told that results would be seen

by others; i.e., that the scale was to be used administratively. A direct comparison of the criterion score distributions for these two groups would be indeterminate, since instructions were consistent within departments, and departmental differences beclouded the issue. The point is important, however, because as discussed earlier, the major reason for forced-choice construction is resistance to bias and leniency. Therefore, a method of maximizing any bias extant should be used in an experiment comparing the graphic versus forced-choice format. Instructions of the two kinds mentioned seem to be the only major kind of class-wide influence likely to be encountered in operational use of student rating of instructors. On this basis, two sets of instructions should be incorporated in the experimental design of a later study.

Summary

The results of this investigation reaffirm the need for research in the social psychology of student rating of teachers. In the development of an instrument useful in such research, the "standard" forced-choice method was modified on various technical grounds and used to construct a validation scale. The scale was administered in graphic form to cross-validate item indexes of biasability and discrimination and as a means of developing scoring keys. Considerable loss of items of high discrimination was experienced in the preparation of the Validation Scale from the Qualification Check List. Still greater loss of items would occur if all technical requirements of forced-choice scale

construction were applied to the construction of a forced-choice scale through use of the cross-validation statistics. In addition, various deficiencies of forced-choice format led to the conclusion that the use of forced-choice format is unsatisfactory for diagnostic purposes. The Validation Scale, scored as a forced-choice form, has high validity in predicting the composite criterion of teacher effectiveness employed. Before conclusions can be drawn regarding the relative effectiveness of forced-choice versus graphic methods, the advantages of various methods of scale construction, and the efficiency of various techniques of controlling bias, further research is necessary. A number of specific research proposals have been presented to clarify the issues involved. The major conclusion of this thesis is that the "standard" method of forced-choice rating scale construction should be modified and further tested before being applied in routine fashion to the construction of a scale to be used by students in rating college instructors.

BIBLIOGRAPHY

1. Barzun, J. Teacher in America. Boston: Little, Brown & Co., 1945.

2. Berkshire, J. R. Investigation of certain assumptions underlying the forced-choice method. Paper read at Amer. Psychol. Ass., San Francisco, Sept., 1955.

3. Brogden, H. E. A rationale for minimum distortion in personality questionnaire keys. Psychometrika, 19, No. 2, June, 1954, 141-148. (See also, Personnel Research Section Report 868, same title, TAG.)

4. Bryan, R. C. Pupil Ratings of Secondary School Teachers. New York: Bureau of Publication, Teachers Coll., Columbia Univer., 1937.

5. Cole, Luella. The Background for College Teaching. New York: Farrar & Rinehart, 1940.

6. Flanagan, J. C. The critical incident technique. Psychol. Bull., 51, No. 4, July, 1954, 327-358.

7. Gilliland, C. E. Study of student rating of faculty practices in member schools. A letter to member deans of the American Association of Collegiate Schools of Business. Personal communication.

8. Harding, F. D. & Long, W. F. Development of a Forced-Choice Rating Scale for Evaluation of Class Room Instructors. Air Training Command Human Resources Research Center, Lackland Air Force Base, San Antonio, Texas, 1952, Research Bulletin 52-21.

9. Henderson, A. D. Evaluation of teacher effectiveness. In F. J. Kelly (Ed.), Improving College Instruction. Washington: American Council on Education Studies, Series 1, No. 48, July, 1951.

10. Highet, G. The Art of Teaching. New York: Alfred A. Knopf, 1950.

11. Highland, R. W. & Berkshire, J. R. A Methodological Study of Forced-Choice Performance Rating. Human Resources Research Center, Lackland Air Force Base, San Antonio, Texas, 1951, Research Bulletin 51-9.

12. Johnson, D. M. Technique for analysis of a highly generalized response pattern. Psychol. Rev., 53, 1946, 348-361.

13. Langmuir, C. R. Cross validation. Test Serv. Bull., The Psychol. Corp., 47, Sept., 1954.

14. Lanman, R. W. & Remmers, H. H. The "preference" and "discrimination" indices in forced-choice scales. Educ. & Psychol. Measmt., 14, No. 3, Autumn 1954, 541-551.

15. Lehman, H. C. Can Students rate teachers? Educ. Administr. & Supervis., 13, October, 1927, 459-466.

16. Loevinger, Jane. Effect of distortions of measurement on item selection. Educ. & Psychol. Measmt., 14, No. 3, Autumn 1954, 441-448.

17. Medalie, R. J. The student looks at college teaching. In T. C. Blegen & R. M. Cooper (Ed.), The Preparation of College Teachers. Washington: American Council on Education, Series 1, No. 14, July, 1950.

18. Morsh, J. E. & Wilder, Eleanor W. Identifying the Effective Instructor: A Review of the Quantitative Studies, 1900-1952. Research Bulletin, Air Force Personnel & Training Research Center, Lackland Air Force Base, San Antonio, Texas. (TR-54-44).

19. O'Hara, J. Appointment with O'Hara. Colliers Magazine, March 16, 1956.

20. President's Commission on Higher Education. Higher Education for American Democracy. New York: Harper & Bros., 1947.

21. Reed, Anna Y. The Effective and Ineffective College Teacher. New York: American Book Co., 1935.

22. Richardson, M. W. Forced-choice performance reports: a modern merit-rating method. Personnel, 26, 1948, 205-212.

23. Riley, J. W., Ryan, B. F., & Lifshitz, Marcia. The Student Looks at His Teacher. An inquiry into the implications of student ratings at the college level. New Brunswick, N. J., Rutgers Univer. Press, 1950.

24. Rundquist, E. A., & Bittner, R. H. A merit rating procedure developed by and for the raters. Personnel, 26, 1950, 273-283.

25. Rundquist, E. A. The forced-choice technique and rating scales. Paper read at Amer. Psychol. Ass., Philadelphia, 1946. Dittoed release.

26. Ryans, D. G. An analysis and comparison of certain techniques for weighting criterion data. Educ. & Psychol. Measmt., 14, No. 3, Autumn 1954, 449-458.

27. Schneider, F. More Than an Academic Question. Berkely, Calif.: The Pestalozzi Press, 1945.

28. Seeley, L. C. Construction of Three Measures for Instructor Evaluation. Technical Report -- SDC 383-1-5, Office of Naval Research, July 20, 1948.

29. Sisson, E. D. Forced-choice -- the new Army rating. Personnel Psychol., 1, 1948, 365-381.

30. Travers, R. M. W. A critical review of the validity and rationale of the forced-choice technique. Psychol. Bull., 48, 1951, 62-70.

31. Woodburne, L. S. Faculty Personnel Policies in Higher Education. New York: Harper & Bros., 1950.

APPENDIX

APPENDIX A

PRELIMINARY ORGANIZATION OF CATEGORIES FOR QUALIFICATION CHECK LIST

No.	Category
1.	Appearance
2.	Attitudes
3.	Classroom Management
4.	Course Content
5.	Diagnostic Abilities
6.	Discipline - Prestige
7.	Discussion
8.	Extra-Class Factors
9.	Global Judgment
10.	Grading and Examination
11.	Laboratory
12.	Lecturing
13.	Personal Stability
14.	Preparation
15.	Rigidity
16.	Scholarship
17.	Social Abilities
18.	Teaching Philosophy
19.	Teaching Techniques
20.	Verbal Abilities

APPENDIX B

CRITICAL INCIDENT QUESTIONNAIRE

DEAR STUDENT:

This a request for your help in a research project to be used to improve college teaching. Your contribution here will be interesting to you and valuable to others.

The procedure used protects both you and the teacher from embarrassment, so you can be completely frank and detailed. Use no names. Do not sign your name.

What actually constitutes "good" and "bad" college teaching? The many past efforts to answer this question have produced results like "good personality," "energetic," "fairminded," etc. These phrases are very hard to use because they mean one thing to one person and something very different to another.

Let's try a new method, the "critical incident technique." This method assumes that characteristics of the superior college teacher are more likely to be determined from analysis of teacher behavior in critical situations rather than in commonplace and everyday situations. What are these "critical situations" in college teaching? In general, any situation which was in any way important to you at the time, is considered "critical."

Here are sample situations which some persons regard as "critical:" The teacher is asked an embarrassing question by the student; a student called upon to recite is obviously unprepared; a student disagrees with the way in which an examination question is graded; etc. Note that these critical situations produce stress or tension in teacher, student, or both.

Your contribution is to describe as many of your experiences in such situations as possible. What were the circumstances? What did the teacher do or fail to do? How did you feel about it then? How do you feel about it now? Even if someone else might consider it trivial or silly, please describe any situation which mattered to you. The incidents may have occurred inside or outside of the class room.

Please ask any questions you wish. If you have none, please turn to the next page.

Page 2

Please think of the college teacher you can judge most accurately (because you have had several, recent or present, courses with this teacher). He is the teacher you know best, whatever the reason. Please describe the incident most unfavorable to him as a teacher. What were the circumstances? What did the teacher do? or fail to do? How did you feel about it then? How do you feel about it now?

(Include other incidents, if possible. Use reverse side if necessary.)

Page 3

Now, using this same teacher, the one you know best, please describe the incident most favorable to him as a teacher. What were the circumstances? What did the teacher do or fail to do? How do you feel about it now?

(Include other incidents, if possible. Use reverse side if necessary.)

Page 4

On this page, please rate the teacher you have just described, the teacher you know best, by checking in one of the boxes.

Worst teacher I have known	Average	Best teacher I have known

Page 5

Now think of the college teacher you liked best of all -- the one you felt most "drawn to," the one who is highest on your list as a person. Even such a teacher occasionally does things you don't like -- things which detract from his teaching effectiveness. Please describe the incident which disappointed you the most. What were the circumstances? What did the teacher do or fail to do? How did you feel about it then? How do you feel about it now?

(If the teacher previously described as "best known" is also "best liked," then describe below the second best liked.)

(Include other incidents, if possible. Use the reverse side if necessary.)

Page 6

Now think of the college teacher with whom you had "little in common," whom you strongly disliked -- perhaps with good reason, perhaps not. But anyway, he is at the bottom of your list as a person. Even in this case you should be able to remember an incident which surprised and pleased you as an example of good teaching. Please describe this incident. What were the circumstances? What did the teacher do? or avoid doing? How did you feel about it then? How do you feel about it now?

(If this least liked teacher should happen to be the best known teacher described earlier, use the teacher second from the bottom on your list.)

(Include other incidents, if possible. Use reverse side if necessary.)

Page 7

I would be glad to have your comments on any phase of college teaching. Thank you very much for your help.

E. B. Cobb, Instructor
Room 21, Psychological Clinic
University of Tennessee

APPENDIX C

QUALIFICATION CHECK LIST (FRONT PAGE)

To show format, Form M of the QCL is appended herewith; complete list of items is presented in Appendix F.

UNIVERSITY INSTRUCTOR RATING RESEARCH ● QUALIFICATION CHECK LIST

A National Study Sponsored by the Department of Psychology
of The University of Tennessee

Orientation and Instructions

ORIENTATION: This is a request for your help in a research project to improve college teaching. Your contribution should be interesting to you and will be valuable to others. *Please read the instructions below carefully* and do the very best you can. The administrator of this check list cannot answer questions, so go ahead and guess if you need to.

The procedure used protects everyone from embarrassment. No names are used. You are safe in being completely frank. If you are not, the results will be misleading.

INSTRUCTIONS: The first thing to do is to select a college teacher out of all those you have known. Later on, you are asked to describe this teacher on a Qualification Check List.

In choosing a teacher, first try to recall "*critical incidents*" in college teaching. These are situations arising in class, lab, shop, office, (or anywhere on or off campus) which produce stress or tension in student, teacher, or both. They interfere with learning. *The teacher may have met a situation very poorly.* For example, a student disagrees in class with the way in which an exam question is graded and the teacher "shuts him up" caustically or ignores him, etc. On the other hand, *the teacher may have handled a situation well.* For example, he is asked an embarrassing question and admits he doesn't know, or looks it up and answers it later, etc.

What class had one or more such incidents? *Reflect a minute now* and choose one particular teacher before going on. Please *do not choose* the teacher in whose class this check list is administered.

Next, describe the teacher you have chosen on the Qualification Check List which follows. The chart below shows the method of indicating how well each descriptive word or phrase fits the teacher you have selected.

Meaning of Letter Ratings and Further Instructions

(VG) G F P VP 6 Circle VG if the word or phrase is a *VERY GOOD description* of this teacher; if he is *the very best teacher you have ever known* in this respect.

VG (G) F P VP 6 Circle G if the word or phrase is a *GOOD description* of this teacher; if his behavior shows very few exceptions from this quality; if this is one of the things you remember about him.

VG G (F) P VP 6 Circle F if the word or phrase is a *FAIR description* of this teacher; if, though variable, the balance is in his favor with respect to this word or phrase.

VG G F (P) VP 6 Circle P if the word or phrase is a *POOR description* of this teacher; if there are somewhat more exceptions than agreements between the phrase and his behavior.

VG G F P (VP) 6 Circle VP if the word or phrase is a *VERY POOR description* of this teacher; if the teacher is *the very worst you can remember* in this respect.

VG G F P VP (6) *Circle 6* if you cannot use the word or phrase; if you *don't know* or it *doesn't apply* at all to this teacher.

QUALIFICATION CHECK LIST

VG G F P VP 6 (1) Has easy yet forceful personality?
VG G F P VP 6 (2) Gives exams that are a fair sampling of content covered?
VG G F P VP 6 (3) Handles emotional or rebellious student so as to maintain respect of class?
VG G F P VP 6 (4) Avoids repeating himself?
VG G F P VP 6 (5) Uses survey quiz or questionnaire at start of course to determine gaps in students' background knowledge?
VG G F P VP 6 (6) Limits lecturing to a minimum during shop period so students can work?
VG G F P VP 6 (7) Guides class well in discussion of controversial subjects?
VG G F P VP 6 (8) Praises more often than he criticizes?
VG G F P VP 6 (9) Gives frequent class reviews?
VG G F P VP 6 (10) Uses sufficient supporting details in lesson?
VG G F P VP 6 (11) Gives exam questions which adequately test student's ability to use the knowledge in the course?
VG G F P VP 6 (12) Unifies the subject in his lectures?

(continued at the top of page)

Form M

Qualification Check List (continued)

VG G F P VP 6 (13) Understands problems most often met by college students in their work?
VG G F P VP 6 (14) Builds respect for points of view other than those held on entering class?
VG G F P VP 6 (15) Provides mimeographed outline of lectures, thus doing away with general note taking?
VG G F P VP 6 (16) Refrains from bluffing?
VG G F P VP 6 (17) Uses evidence from sources other than "paper and pencil tests" to evaluate student progress?
VG G F P VP 6 (18) Avoids duplication with other courses?
VG G F P VP 6 (19) Usually happy rather than gloomy?
VG G F P VP 6 (20) Sets the same standards of attainment for all students?
VG G F P VP 6 (21) Controls class discussions to prevent rambling and confusion?
VG G F P VP 6 (22) Makes an effort to meet the needs of individuals?
VG G F P VP 6 (23) Hard-working?
VG G F P VP 6 (24) Writes legibly on blackboard?
VG G F P VP 6 (25) Stresses accuracy by teaching methods of achieving it?
VG G F P VP 6 (26) Presents material in logical sequence?
VG G F P VP 6 (27) Keeps exam results confidential when returning papers or posting grades?
VG G F P VP 6 (28) Good posture and bearing?
VG G F P VP 6 (29) Shows interest in students as persons?
VG G F P VP 6 (30) Presents controversial questions fairly?
VG G F P VP 6 (31) Has efficient procedure for distributing papers or posting grades?
VG G F P VP 6 (32) Has accurate and up-to-date knowledge of his subject?
VG G F P VP 6 (33) Inspires trust?
VG G F P VP 6 (34) Uses effective figures of speech?
VG G F P VP 6 (35) Uses class average as standard instead of arbitrary standard?
VG G F P VP 6 (36) Provides reasonable lesson objectives?
VG G F P VP 6 (37) Free from annoying personal mannerisms?
VG G F P VP 6 (38) Makes specific daily assignments?
VG G F P VP 6 (39) Addresses students as "Mr.--" or "Miss--" so that they like it?
VG G F P VP 6 (40) Avoids being led into long discussions?
VG G F P VP 6 (41) Gives regular quizzes?
VG G F P VP 6 (42) Avoids appearing inconvenienced by students' requests?
VG G F P VP 6 (43) Accepts text answers even if he disagrees as to correctness, unless issue arose in class previously?
VG G F P VP 6 (44) Shows initiative in using new teaching methods?
VG G F P VP 6 (45) Tactful?
VG G F P VP 6 (46) Develops course along lines of student interests and contributions?
VG G F P VP 6 (47) Uses auditory teaching aids frequently?
VG G F P VP 6 (48) Successfully integrates subject with allied subjects?
VG G F P VP 6 (49) Speaks fluently?
VG G F P VP 6 (50) Fair?
VG G F P VP 6 (51) Has calm, even temperament?
VG G F P VP 6 (52) Interests himself in sports program?
VG G F P VP 6 (53) Finds out why when no one can answer his questions?
VG G F P VP 6 (54) Keeps "poise" and "charm" subordinate to teaching?
VG G F P VP 6 (55) Has much technical knowledge of subject?
VG G F P VP 6 (56) Voice compels attention?
VG G F P VP 6 (57) Criticizes specific acts rather than personalities?
VG G F P VP 6 (58) Organizes course in accordance with catalog description?
VG G F P VP 6 (59) Effectively "draws out" students?
VG G F P VP 6 (60) Has mature appearance?
VG G F P VP 6 (61) Explains difficult material clearly?
VG G F P VP 6 (62) More "human being" than "teacher"?
VG G F P VP 6 (63) Modest but not retiring?
VG G F P VP 6 (64) Avoids heckling students with long and close questioning?
VG G F P VP 6 (65) Patient with students?
VG G F P VP 6 (66) He is impartial in examining and grading, does not discriminate for or against fraternities, athletes, graduate students, women, etc.?

(please continue on reverse side)

APPENDIX C

QUALIFICATION CHECK LIST (BACK PAGE)

To show format, Form M of the QCL is appended herewith; complete list of items is presented in Appendix F.

Further Instructions: Read *all* the phrases in the scale below and *then* check (✓) the *one* box best expressing your opinion of this teacher *as a person.* Please be completely frank. Indicate how you feel about this person rather than how you think you *should* feel. *Do not* turn back to read the Check List before you rate. Rate him as a *person*, not as a *teacher.*

AS A PERSON

FINEST PERSON I EXPECT TO KNOW--WISH I WERE JUST LIKE HIM	*VERY FINE PERSON*--WISH MORE PEOPLE WERE LIKE HIM	*FINE PERSON*--WISH MORE PEOPLE HAD SOME OF HIS QUALITIES	*LIKED HIM VERY MUCH*--BETTER THAN MOST IN MANY WAYS	*ABOUT LIKE MOST PEOPLE*--WE GOT ALONG ALL RIGHT	*HE HAD SOME UNFORTUNATE QUALITIES*--HE ANNOYED ME OCCASIONALLY	*UNFORTUNATE PERSON*--DISLIKED HIM SOMETIMES	*UNPLEASANT PERSON*--I DISLIKED HIM	*EXTREMELY UNPLEASANT*--WORST I EVER KNEW OR EXPECT TO KNOW

Classification Data

This information is vital to the research, so please check each appropriate parenthesis. If you are uncertain, check anyway, making the best guess you can.

This teacher was:

60 years old or over ( )
50-59 years old ( )
40-49 years old ( )
30-39 years old ( )
20-29 years old ( )
19 years old or under ( )

Academic rank was:

Associate Professor or Professor ( )
Instructor or Assistant Professor ( )
Teaching Assistant ( )

Sex:

Male ( )
Female ( )

The effectiveness of a teacher may depend on *what kind* of teaching he does. For example, a teacher *might* be a good lecturer but a poor conference leader, etc. In your experience with this teacher, what was the *main* teaching method used? Put a one (1) by the main teaching method in one of the parentheses below. Next, put in a two (2) for the second most important method. Then continue to rank, putting 3, 4, etc., in the parentheses if any of the other methods listed below had an appreciable influence on your rating of this teacher. Do not rank methods which he did not use at all.

Lecturing ( )
Class discussion ( )
Recitation and oral quizzing ( )
Seminar ( )
Lecture-demonstration (lab. or shop) ( )
Audio-visual demonstration (movies, slides, etc.) ( )
Field work ( )
Individual laboratory work ( )
Shop work ( )
Individual conferences with students (for example, thesis advisor) ( )
Workshop (education) ( )

Were there any *other* opportunities for observation of the teacher which actually affected your description or rating? If so, write them here _____

Your Statistical Description: Please Check (✓). You are:

Sophomore ( )
Junior ( )
Senior ( )
Graduate Student ( )

20 years of age or younger ( )
21-29 years old ( )
30 years or older ( )

NOTE: Graduate students who are teaching should check as Graduate Student.

Male ( )
Female ( )

Important Note: Before returning this form, please check it for omissions. Please do not change any ratings!

A Note of Thanks and Explanation

Thank you very much for your help. This form is one of six, each of which contains different items. The immediate purpose of the research is to find out which items discriminate between good and poor teachers and to what degree. E. B. Cobb

Comments

Further Instructions: Read *all* the phrases in the scale below and *then* check (✓) the *one* box best expressing your judgment of his (her) *effectiveness as a college teacher. Do not* turn back to read the Check List before you rate.

EFFECTIVENESS AS A COLLEGE TEACHER

THE WORST TEACHER I HAVE EVER KNOWN	*DEFINITELY NOT WORTH* HIS SALARY AND MY TIME	*HARDLY WORTH* HIS SALARY AND MY TIME	*JUST WORTH* HIS SALARY AND MY TIME	*DEFINITELY WORTH* HIS SALARY AND MY TIME	*MORE THAN WORTH* HIS SALARY AND MY TIME	*A SUPERIOR* TEACHER	*ONE OF THE TWO OR THREE* BEST TEACHERS I HAVE KNOWN	*THE VERY BEST* TEACHER I HAVE EVER KNOWN

APPENDIX D

DIRECTIONS FOR ADMINISTRATION OF THE QUALIFICATION CHECK LIST

UNIVERSITY INSTRUCTOR RATING RESEARCH -- QUALIFICATION CHECK LIST

E. B. Cobb, The Department of Psychology
The University of Tennessee, Knoxville, Tennessee

Directions for Administration

First steps: Making arrangements.

Arrange the place and hour with the instructor. The attached letter is provided to aid you in securing the instructor's permission for using his class for 30 minutes of his class time. If you wish to use this letter, include a copy of the Qualification Check List. It may also serve as a guide for your use if you wish to phone the instructor instead. Any class in the subject matter area specified in the covering letter will do; very small classes are to be avoided unless they are the only ones possible in the given area. Please do not use Freshman classes. Graduate classes may be included if they have 10 or more students.

Administration must be during the FIRST 20 - 30 minutes of the class hour. Besides being a necessary control, there are administrative advantages if we do not have to contend with the instructor's attempt to introduce the administrator and explain the research. Since such explanation might introduce uncontrolled biases into the situation, the instructor should not be present during administration of the Check List.

There are six forms of the Check List arranged so that by passing them out from the top of the stack downward approximately equal numbers of each are distributed.

Starting:

Since class-room administration is necessary to prevent biased sampling, administration time has been set low (20-30 minutes) in order to gain the participation of the instructor. A short, simple explanation while passing out the Check Lists is necessary to avoid exceeding 30 minutes.

Formal verbatim directions are ineffective in such situations. In your own words cover the following points while passing out lists:

1. Identify yourself.

2. Mr. (Dr.) (Instructor) has agreed to use class time for this research.

3. Read the directions and begin immediately.

4. When everyone is through, class will (a) continue, or (b) be dismissed (depending on prior agreement with instructor).

Avoid jocularities such as reference to "guinea pigs," "a chance to get even," or anything else. Only the points above are to be mentioned.

After starting:

The Qualification Check List has been designed to be self-administrating. Few questions were asked in pilot studies. These are to be handled by the reply: "Read the instructions again and use your best judgment." The orientation paragraph of the Qualification Check List makes this clear.

Standardized administration here calls for complete absence of comments, interpretations, motivating statements, etc.

Dear _____

Colleagues in Psychology at the University of Tennessee have asked me to arrange for the administration of copies of the enclosed Check List to some of our students as part of a national sampling procedure.

As you will note, your own teaching is not to be rated. The Check List is completely anonymous. Scales in existence have defects, and this is an attempt to build better ones.

The research requires the Check List to be administered in class in order to avoid biased sampling. It takes about 20 minutes -- some of the slower students may take 30.

May I come in at the beginning of the hour soon? You could come in at half-time, or I could dismiss the class, as you prefer.

UNIVERSITY INSTRUCTOR RATING RESEARCH -- QUALIFICATION CHECK LIST

E. B. Cobb, The Department of Psychology
The University of Tennessee, Knoxville, Tennessee

ADMINISTRATION LOG

(Please complete, tear off, and return with the Check List in envelope provided)

Course Title _____

Day of Week _____ Date: _____

TIME:	Hour	Minutes
Finish:	_____	_____
Start:	_____	_____
Elapsed:	_____	_____

What was the general response tone of the group?

_____, Administrator
(Signature)

_____
(College or University)

APPENDIX E

GEOGRAPHICAL DISTRIBUTION OF QUALIFICATION CHECK LIST SAMPLE

Region	Total No. of Schools	State	No. of Schools
Far West:	2	California	2
Mountain States:	7	Idaho	1
		Utah	1
		Arizona	1
		Colorado	4
Plains States:	4	Nebraska	1
		Kansas	3
Southwest:	7	Oklahoma	2
		Texas	3
		Arkansas	2
Midwest:	16	Iowa	1
		Missouri	2
		Wisconsin	2
		Illincis	3
		Indiana	1
		Michigan	4
		Ohio	3
Southeast:	13	Mississippi	1
		Alabama	3
		South Carolina	2
		North Carolina	3
		Virginia	3
		West Virginia	1
Middle Atlantic:	16	Maryland	2
		Pennsylvania	4
		New Jersey	2
		New York	8
Northeast:	5	New Hampshire	1
		Massachusetts	3
		Connecticut	1
Total:	70	Total:	70

APPENDIX F

SUMMARY OF QUALIFICATION CHECK LIST RESULTS GROUPED BY ITEM-EFFECTIVENESS CRITERION CORRELATION

QCL Form and Item No.	Item	R_{ie}	R_{ip}	Mean
	0 - .04			
J-57	Uses unannounced quizzes to keep students motivated in their preparation?	.02	.04	3.367
	.05 - .09			
I-52	Avoids use of profanity?	.05	.10	1.640
I-11	Uses objective tests instead of essay tests?	.09	.13	2.787
	.10 - .14			
M-27	Keeps exam results confidential when returning papers or posting grades?	.12	.17	2.413
K-7	Avoids pacing around room?	.12	.15	2.098
I-18	Prevents cheating?	.13	.10	2.113
	.15 - .19			
H-19	Avoids over use or improper use of slang?	.16	.18	1.844
H-9	Leaves room in good condition for next instructor?	.16	.15	1.907
I-65	Generally uses multiple choice and completion-type tests in preference to true-false?	.16	.16	2.590
I-46	Quiet and reserved?	.16	.17	2.614
L-47	Has good hearing?	.16	.12	1.863
K-18	Considers assignments important, holds students to them?	.18	.11	2.259
L-9	Sees that seats and other equipment are properly arranged?	.19	.14	2.443
M-15	Provides mimeographed outline of lectures, thus doing away with general note taking?	.19	.21	3.980
	.20 - .24			
M-38	Makes specific daily assignments?	.21	.18	2.574
J-31	Makes original suggestions, presses them forcibly?	.21	.10	2.881

SUMMARY OF QUALIFICATION CHECK LIST RESULTS GROUPED BY ITEM-EFFECTIVENESS CRITERION CORRELATION (CONTINUED)

QCL Form and Item No.	Item	R_{ie}	R_{ip}	Mean
	.20 - .24 (Continued)			
J-13	Loyal to university policy?	.21	.25	2.243
J-2	Avoids poking fun at authorities or ideas?	.22	.26	2.538
H-13	Avoids shouting?	.23	.32	1.733
I-2	Punctual for appointments?	.24	.20	1.885
J-9	Makes sure all students can see blackboard well?	.24	.17	2.081
J-15	Introduces himself properly at the first session?	.24	.24	2.038
	.25 - .29			
M-47	Uses auditory teaching aids frequently?	.25	.24	3.255
M-31	Has efficient procedure for distributing papers or posting grades?	.25	.29	2.393
J-17	Encourages students to work together?	.25	.31	2.788
L-7	Starts and stops class on time?	.25	.19	2.388
M-43	Accepts text answers even if he disagrees as to correctness, unless issue arose in class previously?	.25	.30	2.871
M-24	Writes legibly on blackboard?	.26	.20	2.348
I-4	Makes regular assignments?	.26	.19	2.150
I-60	Refuses to change course grade to placate student or because of other pressures?	.27	.25	2.187
M-41	Gives regular quizzes?	.27	.44	2.728
M-39	Addresses students as "Mr.--" or "Miss" so that they like it?	.27	.35	2.323
M-28	Good posture and bearing?	.27	.36	2.051
M-20	Sets the same standards of attainment for all students?	.27	.27	2.337
L-14	Has good vision?	.27	.19	2.075
M-18	Avoids duplication with other courses?	.28	.26	2.257
M-5	Uses survey quiz or questionnaire at start of course to determine gaps in students' background knowledge?	.28	.27	3.798
J-27	Avoids using class time discussing his convictions on controversial problems irrelevant to course?	.28	.28	2.272
I-7	Good physique?	.28	.35	2.647
L-2	Dresses appropriately and in good taste?	.29	.31	2.177

SUMMARY OF QUALIFICATION CHECK LIST RESULTS GROUPED BY ITEM-EFFECTIVENESS CRITERION CORRELATION (CONTINUED)

QCL Form and Item No.	Item	R_{ie}	R_{ip}	Mean
	.25 - .29 (Continued)			
L-55	Checks and controls light, heat, ventilation, etc., well?	.29	.23	2.607
K-3	Spends minimum of time on roll call?	.29	.26	1.679
I-31	Grades students on performance, not on class attendance?	.29	.33	2.349
J-54	Moral conduct excellent?	.29	.31	1.658
	.30 - .34			
I-39	Plans class trips and expeditions to secure effective learning?	.31	.23	3.428
H-55	Avoids talking to the wall or blackboard?	.31	.37	1.811
M-35	Uses class average as standard instead of arbitrary standard?	.31	.38	2.644
I-13	Has business-like attitude?	.31	.24	2.134
L-51	Explains relationship between absences and grades?	.31	.28	2.589
I-20	Gives enough time on exams so that the average student can complete them?	.32	.30	2.282
M-40	Avoids being led into long discussions?	.32	.23	2.645
K-26	Dresses neatly?	.32	.33	1.958
K-8	Gives student enough time to think when asking questions?	.32	.40	2.355
M-62	Interests himself in sports program?	.33	.33	2.855
K-1	Prompt?	.33	.23	1.980
H-23	Avoids favoring students of opposite sex?	.33	.31	1.942
H-25	Keeps temper?	.33	.48	2.045
L-43	Avoids being sidetracked into discussion not pertinent to lesson?	.33	.16	2.579
L-18	Uses visual aids frequently?	.33	.28	2.970
L-5	Dispenses with unnecessary formality?	.33	.41	2.072
L-28	Avoids becoming so involved in discussion with one student that he forgets rest of class?	.34	.29	2.293
K-61	Refrains from talking about himself?	.34	.31	2.266
I-53	Makes certain that instructional materials (texts, library books, laboratory equipment, etc.) are available to students?	.34	.31	1.946
J-47	Avoids "lecturing" on cutting classes?	.34	.34	2.077

SUMMARY OF QUALIFICATION CHECK LIST RESULTS GROUPED BY ITEM-EFFECTIVENESS CRITERION CORRELATION (CONTINUED)

QCL Form and Item No.	Item	R_{ie}	R_{ip}	Mean
	.30 - .34 (Continued)			
H-17	Avoids embarrassing or ridiculing a student in the presence of others?	.34	.52	2.504
H-24	Firm?	.34	.16	2.193
H-43	Calls students by first name so that they like it?	.34	.51	3.331
	.35 - .39			
J-40	Protects students from danger by organizing and enforcing proper routines in laboratory?	.35	.39	2.259
H-51	Has necessary equipment ready in the classroom?	.35	.27	2.000
J-66	Conducts classes informally?	.35	.39	2.059
J-64	Frank?	.35	.32	1.806
I-64	Avoids taking too much time making assignments?	.35	.30	1.860
I-58	Provides for, announces, and maintains reasonable office hours?	.35	.30	2.193
I-40	Free from pecularities in appearance?	.35	.38	2.228
I-9	Is especially effective in campus affairs?	.35	.41	2.817
K-20	Reports exam results promptly?	.35	.32	2.448
H-45	Permits making up quizzes or exams when absence is explained?	.36	.43	2.134
H-48	Follows the course outline?	.36	.25	2.370
H-27	Gives examinations and quizzes neither too often nor too seldom?	.36	.37	2.322
K-56	Prevents "horse play" in lab and shop courses?	.36	.21	1.983
I-12	Voice carries to everybody in room?	.36	.29	1.678
J-24	Requires the proper amount of work for the credit granted?	.36	.32	2.224
L-8	Has had practical experience in his field?	.36	.31	1.714
H-44	Furnishes a well-prepared, written course outline?	.37	.28	3.059
H-65	Minds his own business?	.37	.43	2.158
M-58	Organizes course in accordance with catalog description?	.37	.35	2.307
I-54	Maintains up-to-date bibliographies?	.37	.26	2.233
J-18	Refers to past learning when teaching new lessons?	.37	.33	2.044

SUMMARY OF QUALIFICATION CHECK LIST RESULTS GROUPED BY ITEM-EFFECTIVENESS CRITERION CORRELATION (CONTINUED)

QCL Form and Item No.	Item	R_{ie}	R_{ip}	Mean
	.35 - .39 (Continued)			
L-48	Finds it unnecessary to use force to keep order?	.37	.35	1.805
M-64	Avoids heckling students with long and close questioning?	.38	.51	2.272
I-14	Encourages classroom discussion among students instead of limiting it to teacher-student conversations?	.38	.40	2.361
J-39	Has good educational background?	.38	.30	1.474
J-7	Avoids speaking too rapidly or too slowly?	.38	.33	2.339
I-6	Has students apply new material right after presentation?	.38	.27	2.658
K-36	Avoids erasing useful material from the blackboard too soon?	.38	.35	2.069
K-30	Avoids poking fun at people?	.38	.47	2.215
L-61	Holds conference with each student?	.38	.42	2.992
I-29	Optimistic?	.39	.46	2.358
J-5	Seldom shows "It's right or completely wrong" attitude?	.39	.46	2.461
J-35	Removes excessive emotional pressure on student in discussion by shifting to another student?	.39	.52	2.595
J-53	Uses novel or startling statements to start discussion?	.39	.36	2.990
H-12	Avoids taking too much time drawing or writing on blackboard?	.39	.29	2.001
	.40 - .44			
M-6	Limits lecturing to a minimum during shop period so students can work? (R_{ip} not available, IBM error, must go back to original data.)	.40		2.610
H-32	Offers his services to all and gives his phone number and address?	.40	.46	2.659
M-8	Praises more often than he criticises?	.40	.50	2.805
J-32	Summarizes at the end of the period?	.40	.45	3.365
J-16	Attacks problems in terms of principles or issues rather than personalities?	.40	.40	2.053
K-19	Avoids wandering from topics under discussion?	.40	.27	2.499
M-4	Avoids repeating himself?	.41	.36	2.543

SUMMARY OF QUALIFICATION CHECK LIST RESULTS GROUPED BY ITEM-EFFECTIVENESS CRITERION CORRELATION (CONTINUED)

QCL Form and Item No.	Item	R_{ie}	R_{ip}	Mean
	.40 - .44 (Continued)			
H-66	Pronounces words clearly?	.41	.34	1.853
H-29	Encourages students to search for reading material?	.41	.32	2.744
J-22	Improvises training aids when necessary?	.41	.61	2.713
L-58	Explains his position but does not argue with students?	.42	.49	2.446
K-54	Avoids permitting some students to monopolize class discussion?	.42	.37	2.457
I-26	Spends enough time "getting acquainted" with students?	.42	.51	2.886
H-34	Courteous?	.42	.61	2.064
J-28	Finds out if a particular student's viewpoint represents that of class before discussing it at length?	.42	.42	3.012
I-32	Allows students to aid in developing solutions to problems?	.42	.43	2.344
M-60	Has mature appearance?	.42	.43	1.821
L-45	Contributes his free time generously to student activities?	.42	.54	2.661
L-15	Stresses accuracy by illustrating the sources of inaccuracy?	.42	.30	2.289
I-57	Makes effective use of gestures?	.43	.37	2.496
H-16	Good memory?	.43	.29	1.841
H-14	Impartial in his treatment of students?	.43	.52	2.393
J-62	Maintains helpful attitude even when student is unprepared?	.43	.57	2.767
J-19	Explains new terms and writes them on blackboard?	.43	.34	2.155
I-10	Interrupts students tactfully in order to save class time?	.43	.44	2.654
K-64	Avoids being "too strict"?	.43	.53	2.334
L-60	Never appears annoyed when approached outside of class?	.43	.56	2.014
K-32	Lets students know in advance the kind of tests to be given?	.44	.40	2.359
L-37	Sincerely polite to students outside of class?	.44	.62	1.876

SUMMARY OF QUALIFICATION CHECK LIST RESULTS GROUPED BY ITEM-EFFECTIVENESS CRITERION CORRELATION (CONTINUED)

QCL Form and Item No.	Item	R_{ie}	R_{ip}	Mean
	.40 - .44 (Continued)			
K-47	Impresses people favorably on first acquaintance?	.44	.51	2.396
I-5	Avoids being stern or dominating?	.44	.59	2.353
I-25	Assigns questions for next class period when discussion does not clarify them?	.44	.35	3.071
I-46	Encourages group activity and initiative?	.44	.40	2.750
F-9	Gives frequent class reviews?	.44	.36	2.872
F-17	Uses evidence from sources other than "paper and pencil tests" to evaluate student progress	.44	.38	2.702
I-37	Uses words understandable to students or discovers which ones are not?	.44	.38	2.080
K-55	Has much technical knowledge of subject?	.44	.39	1.562
H-18	Frank and straightforward about his expectations of the class?	.44	.35	2.135
H-41	Creates student participation by asking questions?	.44	.39	2.455
H-59	Controls class discussions so as to prevent individuals from being hurt?	.44	.57	2.460
H-62	Avoids using threat of poor grades?	.44	.56	2.202
	.45 - .49			
M-66	He is impartial in examining and grading, does not discriminate for or against fraternities, athletes, graduate students, women, etc.?	.45	.43	1.903
J-58	Has good manners?	.45	.53	1.939
J-10	Gives explicit instructions?	.45	.31	2.360
K-48	Distressed rather than complacent at student failure?	.45	.52	2.803
K-45	Makes assignments to strengthen students' weak points?	.45	.42	2.907
K-34	Recognizes and greets students outside of class?	.45	.50	2.042
K-22	Encourages students to express themselves?	.45	.48	2.360
K-17	Usually relaxed rather than tense?	.45	.48	1.958
L-58	Keeps students informed regarding progress in the course?	.45	.39	2.753

SUMMARY OF QUALIFICATION CHECK LIST RESULTS GROUPED BY ITEM-EFFECTIVENESS CRITERION CORRELATION (CONTINUED)

QCL Form and Item No.	Item	R_{ie}	R_{ip}	Mean
	.45 - .49 (Continued)			
L-17	Emphasizes importance of key materials by offering drill or practice exercises?	.45	.32	2.716
L-4	Uses idioms well?	.45	.35	2.282
M-51	Has calm, even temperament?	.46	.60	2.203
H-40	Emphasizes and writes key points on blackboard?	.46	.34	2.332
J-21	Realizes that students learn only after practice?	.46	.43	2.357
K-44	Interrupts a student answering a question only when necessary to clarify or summarize?	.46	.51	2.224
L-50	Discusses questions he raises when students cannot?	.46	.42	2.105
I-61	Avoids overworking certain words?	.47	.37	2.282
I-62	Conducts himself so as not to appear ineffectual or easily "walked over?"	.47	.27	1.880
H-20	Friendly?	.47	.65	2.104
H-33	Avoids trying to be impressive?	.47	.60	2.501
J-55	Points out source of errors on exam questions so students understand mistakes?	.47	.48	2.378
J-52	Admits the difficulty of some content and makes allowance for imperfect mastery?	.47	.55	2.516
J-63	Avoids overworking certain ideas?	.47	.55	2.666
J-49	Willing to help students with research?	.47	.58	2.086
J-42	Associates with students without becoming too familiar or intimate?	.47	.55	2.246
J-34	Sets high standards for self and students?	.47	.35	2.051
J-14	Avoids being "bossy?"	.47	.62	2.529
J-6	Encourages students to ask for help?	.47	.33	2.374
H-50	Makes students feel at ease in class?	.47	.64	2.421
L-20	Talks to everyone including those in back row?	.47	.45	1.963
I-27	Keeps promises?	.48	.45	2.033
I-50	Helps students with their personal problems?	.48	.54	2.828
M-49	Speaks fluently?	.48	.43	1.953
H-61	Has effective vocabulary?	.48	.40	1.743
H-49	Avoids using routine memory-type quizzes?	.48	.42	2.539
H-52	Tolerant?	.48	.65	2.400
H-57	Uses tests to help students review and organize subject matter?	.48	.47	2.906

SUMMARY OF QUALIFICATION CHECK LIST RESULTS GROUPED BY ITEM-EFFECTIVENESS CRITERION CORRELATION (CONTINUED)

QCL Form and Item No.	Item	R_{ie}	R_{ip}	Mean
	.45 - .49 (Continued)			
J-36	Explains basis of grading, helping students prepare better thereafter?	.48	.47	2.512
K-27	Uses clear, simple language?	.48	.45	1.956
L-21	Assigns useful and relevant supplementary materials?	.49	.35	2.414
L-11	Helps students see that good ideas seldom work perfectly?	.49	.44	2.499
L-30	Devoted to teaching?	.49	.40	2.170
L-41	Replies so as to promote productive discussion rather than to cut it off?	.49	.50	2.567
K-33	Expects suggestions; but knows it is his business to direct activities?	.49	.51	2.442
K-51	Provides adequate hints for studying outside reading assignments?	.49	.46	2.729
K-58	Reflects high social ideals in his behavior?	.49	.49	2.021
I-30	Willingly explains material of earlier courses when requested to do so?	.49	.54	2.336
H-58	Pleasing personal appearance?	.49	.47	2.059
M-16	Refrains from bluffing?	.49	.53	2.111
M-63	Modest but not retiring?	.49	.59	2.586
	.50 - .54			
H-60	Ambitious?	.50	.36	2.011
H-30	Corrects in class errors he has made previously?	.50	.52	2.236
M-57	Criticises specific acts rather than personalities?	.50	.55	2.284
J-61	Bases opinions on evidence wherever possible rather than authority?	.50	.49	2.110
J-50	Avoids talking down to students?	.50	.63	2.300
J-26	Distributes time appropriately over outline or topics of the course?	.50	.45	2.469
J-8	His opinions on his subject are dependable?	.50	.42	1.995
I-22	Separates fact from opinion in discussion?	.50	.44	2.220
K-60	Patient in discussion with a student who is wrong but insists he is right?	.50	.58	2.642

SUMMARY OF QUALIFICATION CHECK LIST RESULTS GROUPED BY ITEM-EFFECTIVENESS CRITERION CORRELATION (CONTINUED)

QCL Form and Item No.	Item	R_{ie}	R_{ip}	Mean
	.50 - .54 (Continued)			
K-52	Uses eye-catching, attractive training aids?	.50	.45	3.131
K-53	Organizes instruction into problems?	.50	.42	2.759
K-42	Grades exam items on basis of own lectures, not departmental answers?	.50	.40	2.054
K-39	Avoids including much minor detail in examination?	.50	.45	2.725
K-40	Patient with students who have trouble with basic concepts of course?	.50	.59	2.565
L-26	Avoids covering up lack of knowledge?	.50	.51	2.218
L-19	Welcomes differences of opinion?	.50	.57	2.596
L-6	Aware that other people desire recognition and credit?	.50	.61	2.441
I-36	Sees that examination content is narrow enough so that preparation is possible?	.51	.45	2.538
H-10	Provides a good proportion between lecture and discussion?	.51	.47	2.549
H-31	Bases grades on number and type of errors instead of merely on final answers?	.51	.46	2.667
M-19	Usually happy rather than gloomy?	.51	.62	2.001
M-54	Keeps "poise" and "charm" subordinate to teaching?	.51	.54	2.399
J-60	Allows students opportunity to express themselves?	.51	.52	2.176
J-30	Leads students to take responsibility in planning and checking their own progress?	.51	.44	2.621
J-25	Encourages original thinking?	.51	.47	2.302
K-46	Teaches how to apply the method of science?	.51	.40	2.589
K-28	Carefully fixes responsibility before taking action?	.51	.52	2.418
K-23	Well-bred?	.51	.54	1.843
K-14	Tries to give special help to students having obvious difficulty with course?	.51	.55	2.630
L-57	At ease in class?	.51	.51	1.767
L-49	Makes assignments to fit students' particular interests?	.51	.51	3.000

SUMMARY OF QUALIFICATION CHECK LIST RESULTS GROUPED BY ITEM-EFFECTIVENESS CRITERION CORRELATION (CONTINUED)

QCL Form and Item No.	Item	R_{ie}	R_{ip}	Mean
	.50 - .54 (Continued)			
L-40	Uses examinations to test knowledge of subject matter rather than academic skills?	.51	.45	2.309
J-23	Has pleasant voice?	.52	.51	2.437
M-32	Has accurate and up-to-date knowledge of his subject?	.52	.52	1.576
M-21	Controls class discussions to prevent rambling and confusion?	.52	.43	2.293
M-23	Hard-working?	.52	.46	1.964
J-12	Recognizes and rewards effort by helpful comments?	.52	.57	2.603
K-50	Uses examples, stories, and experiences to show practical value of lesson?	.52	.51	2.000
K-55	Willing to admit errors in grading and change grades on the records?	.52	.55	2.456
J-43	Lets students know things to stress in studying?	.52	.48	2.572
L-33	Avoids wasting time on superfluous explanation?	.52	.35	2.634
I-23	Emotionally mature and stable?	.53	.55	1.878
H-64	Recognizes and allows for variations in intelligence and background among students?	.53	.59	2.780
L-44	Plans lab work that adequately supplements class work?	.53	.40	2.450
H-7	Willing to change teaching methods in response to class needs?	.53	.61	2.787
M-42	Avoids appearing inconvenienced by students' requests?	.53	.59	2.386
K-35	Emphasizes significant points with his voice?	.53	.43	2.330
L-46	Meets difficulties with poise?	.53	.52	2.314
M-37	Free from annoying personal mannerisms?	.53	.60	2.497
M-36	Provides reasonable lesson objectives?	.53	.56	2.422
I-28	Exercises judgment in grading instead of using fixed percentage of A's, B's, C's, etc. in every class?	.53	.59	2.485
I-35	Avoids belittling students' questions?	.53	.59	2.348
M-48	Successfully integrates subject with allied subjects?	.53	.46	2.373

SUMMARY OF QUALIFICATION CHECK LIST RESULTS GROUPED BY ITEM-EFFECTIVENESS CRITERION CORRELATION (CONTINUED)

QCL Form and Item No.	Item	R_{ie}	R_{ip}	Mean
	.50 - .54 (Continued)			
K-29	Observant?	.53	.46	2.309
L-62	Sincere?	.53	.64	1.881
L-38	Reliable and dependable?	.53	.49	1.914
L-16	Presents essential features clearly in blackboard diagrams?	.54	.47	2.409
L-10	Able to stop lecturing to respond appropriately to questions or comments?	.54	.53	2.130
I-34	Shoulders heavy work loads without complaint?	.54	.48	2.153
I-38	Alert?	.54	.47	1.783
H-8	Eliminates ambiguous exam questions in assigning grades?	.54	.50	2.906
L-1	Sticks to a job until it's finished?	.54	.42	2.044
L-13	Enthusiastic and interested in his subject?	.54	.44	1.631
L-54	Fulfills obligations regularly?	.54	.46	2.129
I-16	Easy to get along with?	.54	.71	2.154
I-19	Varies speed of speech to suit meaning?	.54	.42	2.451
I-59	Avoids personal reaction to disagreement with his opinion?	.54	.55	2.471
J-4	Respects the rights of other people?	.54	.66	2.189
J-33	Does not waste time or words?	.54	.43	2.650
M-25	Stresses accuracy by teaching methods of achieving it?	.54	.40	2.560
M-53	Finds out why when no one can answer his questions?	.54	.54	2.876
H-15	Lectures well without depending on notes?	.54	.33	2.102
	.55 - .59			
H-54	Makes allowances for students' mistakes?	.55	.63	2.766
H-6	Dignified but not stuffy?	.55	.56	2.230
M-30	Presents controversial questions fairly?	.55	.60	2.351
J-48	Respects personal confidences?	.55	.63	2.055
J-45	Conducts classes in an assured, confident manner?	.55	.43	1.880
J-44	Gives clear exam questions?	.55	.48	2.501

SUMMARY OF QUALIFICATION CHECK LIST RESULTS GROUPED BY ITEM-EFFECTIVENESS CRITERION CORRELATION (CONTINUED)

QCL Form and Item No.	Item	R_{ie}	R_{ip}	Mean
	.55 - .59 (Continued)			
J-20	Presents a good balance between theoretical and practical knowledge?	.55	.48	2.213
I-47	Reviews and summarizes frequently throughout the course?	.55	.49	2.703
I-21	Criticises so that students do not resent it?	.55	.61	2.460
K-38	Constantly improves his knowledge of subject through continued study and research?	.55	.45	1.734
L-31	Presents various aspects of controversial issues and encourages students to form own opinions?	.55	.34	2.636
H-36	Offers many opportunities to apply learning through activities?	.56	.49	2.906
M-1	Has easy yet forceful personality?	.56	.67	2.288
M-26	Presents material in logical sequence?	.56	.49	2.142
K-49	When necessary, reformulates or rephrases questions he asks students?	.56	.56	2.096
H-56	Discusses with students their difficulties and failures?	.56	.59	2.608
K-37	Treats students as ladies and gentlemen?	.56	.61	1.971
K-15	Uses test results to help find weak points in his teaching?	.56	.54	3.087
L-42	Has pleasant expression?	.56	.65	2.226
I-17	Listens to and respects student opinion regarding exam questions?	.57	.62	2.601
H-63	Responsive to students' expressions?	.57	.64	2.461
M-65	Patient with students?	.57	.70	2.373
J-51	Open Minded?	.57	.70	2.319
J-1	Shows usefulness of the course content?	.57	.49	2.152
I-43	Inspires confidence?	.57	.70	2.589
H-3	Respects opinions of students?	.57	.66	2.481
K-37	Uses carefully constructed and specific essay questions?	.57	.47	2.741
I-44	Notes and explains disagreement between text and lecture?	.57	.50	2.235
I-48	Makes proper use of qualifying statements?	.57	.48	2.349
I-63	Explains all diagrams fully?	.57	.49	2.162

SUMMARY OF QUALIFICATION CHECK LIST RESULTS GROUPED BY ITEM-EFFECTIVENESS CRITERION CORRELATION (CONTINUED)

QCL Form and Item No.	Item	R_{ie}	R_{ip}	Mean
	.55 - .59 (Continued)			
H-28	Assigns significant and interesting research problems?	.57	.50	2.884
M-14	Builds respect for points of view other than those held on entering class?	.57	.62	2.421
K-57	Maintains energy throughout class periods?	.57	.49	1.995
J-38	Accepts and discusses student disagreement rather than ignoring it?	.57	.61	2.261
L-64	Uses class discussion to find student difficulties with new material?	.57	.52	2.741
H-38	Controls class discussions by interpreting what has been said	.58	.48	2.461
M-62	More "human being" than "teacher?"	.58	.72	2.459
K-9	Has high sense of personal integrity?	.58	.60	1.931
L-53	Reasonably quick in getting point that student makes?	.58	.49	2.223
M-45	Tactful?	.58	.49	2.500
K-11	Gives fair examination questions?	.58	.50	2.459
H-2	Stresses aims of course rather than grades?	.58	.43	2.181
K-63	Welcomes conferences and through them provides effective help?	.58	.60	2.386
J-41	Apportions the proper weight to tests, reports, recitations, and final exams in grading?	.58	.52	2.365
L-27	Shows confidence in students?	.58	.64	2.502
L-32	Realizes he is working first with people and second with subject matter?	.59	.62	2.714
K-25	Uses tests to instruct as well as grade?	.59	.56	2.759
J-46	Accepts criticism and acts on it?	.59	.66	2.686
I-33	Pleasant?	.59	.72	2.059
M-44	Shows initiative in using new teaching methods?	.59	.55	2.626
M-46	Develops course along lines of student interests and contributions?	.59	.62	2.591
M-50	Fair?	.59	.67	2.220
	.60 - .64			
H-26	Understands students' point of view?	.60	.70	2.617

SUMMARY OF QUALIFICATION CHECK LIST RESULTS GROUPED BY ITEM-EFFECTIVENESS CRITERION CORRELATION (CONTINUED)

QCL Form and Item No.	Item	R_{ie}	R_{ip}	Mean
	.60 - .64 (Continued)			
M-59	Effectively "draws out" students?	.60	.59	2.674
M-56	Voice compels attention?	.60	.49	2.385
M-10	Uses sufficient supporting details in lesson?	.60	.43	2.078
M-2	Gives exams that are a fair sampling of content covered?	.60	.52	2.321
M-3	Handles emotional or rebellious student so as to maintain respect of class?	.60	.67	2.391
J-65	Commands such respect that students do not cheat?	.60	.64	2.360
M-41	Judges group interests well?	.60	.65	2.505
L-25	Enlivens class with materials not taken from readings?	.60	.53	2.306
L-22	Has ability to think "on the spot?"	.60	.45	2.000
L-36	Uses relevant supporting details?	.61	.45	2.249
M-29	Shows interest in students as persons?	.61	.71	2.336
I-51	Has well organized plans for the course?	.61	.47	2.077
H-5	Takes charge and leads class without lost motion?	.61	.41	2.319
K-66	Exhibits good sportsmanship?	.61	.72	2.191
K-21	Admits mistakes and does not offer excuses?	.61	.62	2.373
L-29	Clearly explains major objectives throughout the course?	.61	.50	2.428
L-12	Grades on the basis of knowledge shown rather than agreement with his own opinions?	.61	.56	2.455
H-39	Uses appropriate, well integrated course materials (texts, lectures, labs, quizzes, and supplementary materials)?	.62	.48	2.404
M-22	Makes an effort to meet the needs of individuals?	.62	.64	2.604
J-37	Makes clear transitions from one point to another by showing relationship?	.62	.52	2.298
I-3	Willingly explains a point in more detail upon request?	.62	.61	1.983
K-6	Gives necessary subject background as basis of discussion?	.62	.53	2.222
L-52	Has spent considerable time in research and study for each lecture?	.62	.41	2.266

SUMMARY OF QUALIFICATION CHECK LIST RESULTS GROUPED BY ITEM-EFFECTIVENESS CRITERION CORRELATION (CONTINUED)

QCL Form and Item No.	Item	R_{ie}	R_{ip}	Mean
	.60 - .64 (Continued)			
K-10	Gives carefully prepared class reviews?	.63	.50	2.823
H-47	Helps students see inconsistencies in ideas?	.63	.58	2.496
M-34	Uses effective figures of speech?	.63	.60	2.036
H-4	Ends lectures effectively?	.63	.50	2.766
M-11	Gives exam questions which adequately test student's ability to use the knowledge of the course?	.63	.54	2.354
M-12	Unifies the subject in his lectures?	.63	.57	2.300
J-11	Creates loyalty among students?	.63	.67	2.770
L-24	Gets cooperation of class in setting up specific goals?	.63	.63	2.782
L-63	Lectures so that students can take good notes?	.64	.46	2.555
L-65	Flexible in adapting to changing needs?	.64	.58	2.494
M-13	Understands problems most often met by college students in their work?	.64	.67	2.379
I-49	Teaches so as to stimulate interest in additional work in the field of study?	.64	.53	2.628
H-35	Able to explain a low grade in conference so as to help student?	.64	.63	2.491
	.65 - .69			
H-22	Encourages initiative on part of students?	.65	.55	2.654
H-1	Inspires you to make the maximum preparation for each day's assignment in his class?	.65	.47	2.755
M-33	Inspires trust?	.65	.69	2.224
J-59	Analyzes class difficulties thoroughly?	.65	.61	2.718
J-56	Has mature and balanced judgment?	.65	.70	1.990
J-29	Conscientious and thorough?	.65	.55	2.020
I-42	Teaches so that students' out-of-class interest is aroused?	.65	.57	2.649
K-65	Uses carefully prepared introduction?	.65	.47	2.611
K-59	Prepares summary carefully?	.65	.51	2.605
K-12	Makes effective use of a pleasing sense of humor?	.65	.69	2.441
L-34	Makes appropriate number of major points in lesson?	.65	.47	2.421

SUMMARY OF QUALIFICATION CHECK LIST RESULTS GROUPED BY ITEM-EFFECTIVENESS CRITERION CORRELATION (CONTINUED)

QCL Form and Item No.	Item	R_{ie}	R_{ip}	Mean
	.65 - .69 (Continued)			
L-23	Makes his quizzes and exams agree with course objectives and content?	.65	.53	2.207
I-41	Teaches at his best consistently?	.66	.57	2.176
M-7	Guides class well in discussion of controversial subjects?	.66	.63	2.435
K-5	Enthusiastic in his teaching?	.66	.49	1.951
K-4	Gives concise instructions?	.66	.55	2.431
K-2	Uses apt and concrete examples of principles?	.66	.53	2.126
L-39	Uses questions skillfully to clarify or define problems under discussion?	.66	.56	2.501
H-11	Has many novel and stimulating ideas and viewpoints?	.67	.58	2.418
I-55	Answers student questions directly without evasion?	.68	.59	2.208
I-56	Inspires a strong atmosphere of mutual good will in his class?	.69	.71	2.472
I-8	Makes sure students understand difficult points?	.69	.62	2.366
K-16	Secures wholehearted cooperation of class?	.69	.67	2.675
I-1	Paces lectures properly in speed and content to students' comprehension?	.69	.56	2.341
I-24	Holds the interest of more than just the brightest pupils?	.69	.58	2.350
J-3	Impresses people favorably after they know him well?	.69	.74	2.261
H-53	Makes assignments challenging and valuable?	.69	.46	2.831
	.70 - .74			
L-59	Directs discussion skillfully?	.70	.58	2.459
H-46	Interprets abstract ideas and theories clearly?	.71	.60	2.464
I-15	Suggestions and advice are extremely valuable?	.71	.62	2.387
K-31	Scholarly but not dull?	.71	.61	2.215
K-43	Meets definite objectives of course?	.72	.52	2.165
K-24	Makes major points clear?	.72	.57	2.085
K-62	Gives well organized lectures?	.73	.54	2.418

SUMMARY OF QUALIFICATION CHECK LIST RESULTS GROUPED BY ITEM-EFFECTIVENESS CRITERION CORRELATION (CONTINUED)

QCL Form and Item No.	Item	R_{ie}	R_{ip}	Mean
	.70 - .74 (Continued)			
M-61	Explains difficult material clearly?	.73	.73	2.290
H-42	Uses vivid supporting details in lectures?	.73	.45	2.392
L-35	Analyzes problems clearly?	.73	.55	2.328
K-13	Presents a subject so that it is fresh and vital?	.74	.63	2.469
H-21	Makes laboratory work meaningful rather than routine exercises?	.74	.58	2.550
	.75 - .79			
L-3	Able to impart his enthusiasm to students?	.75	.60	2.481

APPENDIX G

VALIDATION SCALE (PAGE 1)

RATING SCALE FOR COLLEGE TEACHERS

Teacher Code Number_____. College or University_____

Course Title_____. Student's year in college (circle)
1 2 3 4 grad.

DIRECTIONS

Please describe or rate the teacher in relation to each item of the scale below as accurately and honestly as possible by putting an "X" in one of the boxes. Work rapidly but not thoughtlessly or carelessly. Do not let anyone else influence your judgment. Do not omit items--guess when you must. Do not sign your name.

MEANING OF LETTER RATINGS

O means outstanding; the teacher is the very best you have ever known in this respect.

G means good; the teacher is well above the average of the teachers you have known.

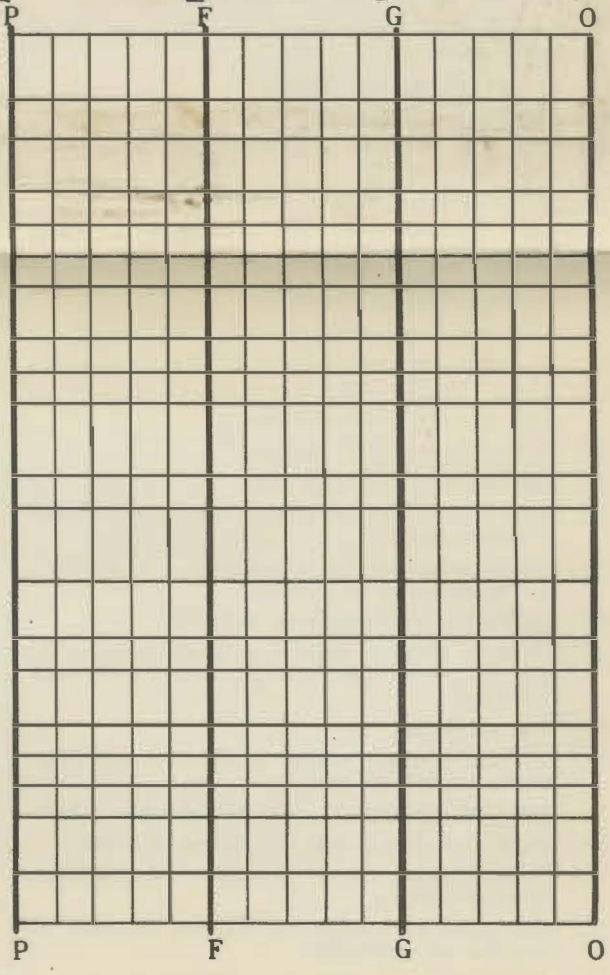
F means fair; the teacher generally behaves acceptably here; the balance is slightly in his favor in this respect.

P means poor; the teacher is the very worst you have ever known in this respect.

IMPORTANT

Check only within boxes; not on vertical lines.

1. Good posture and bearing.
2. Uses sufficient supporting details in lesson.
3. Scholarly but not dull.
4. Has pleasant expression.
5. Makes regular assignments.
6. Frank and straightforward about his expectations of the class.
7. Ambitious.
8. Dresses neatly.
9. Avoids using class time discussing his convictions on controversial problems irrelevant to course.
10. Avoids overworking certain words.
11. Gives exam questions which adequately test student's ability to use the knowledge in the course.
12. Willingly explains material of earlier courses when requested to do so.
13. Free from annoying personal mannerisms.
14. Uses questions skillfully to clarify or define problems under discussion.
15. Stresses aims of course rather than grades.
16. Minds his own business.
17. Free from pecularities in appearance.
18. Reasonably quick in getting point that student makes.
19. Suggestions and advice are extremely valuable.

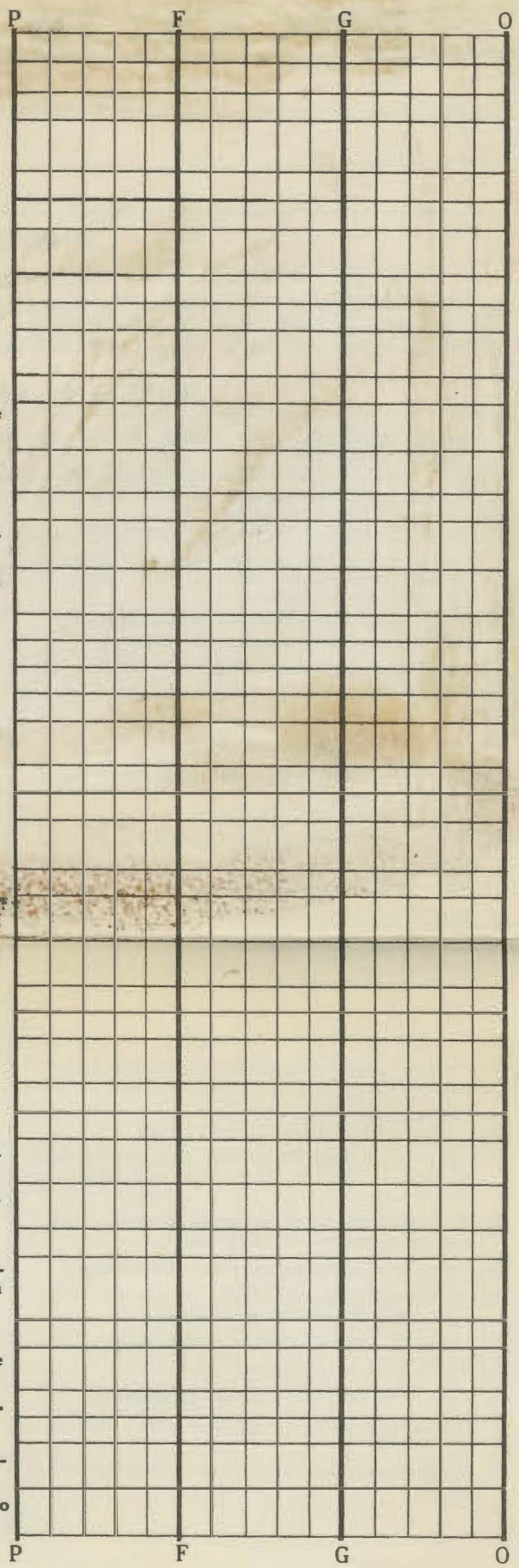


APPENDIX G

VALIDATION SCALE (PAGE 2)

P F G O

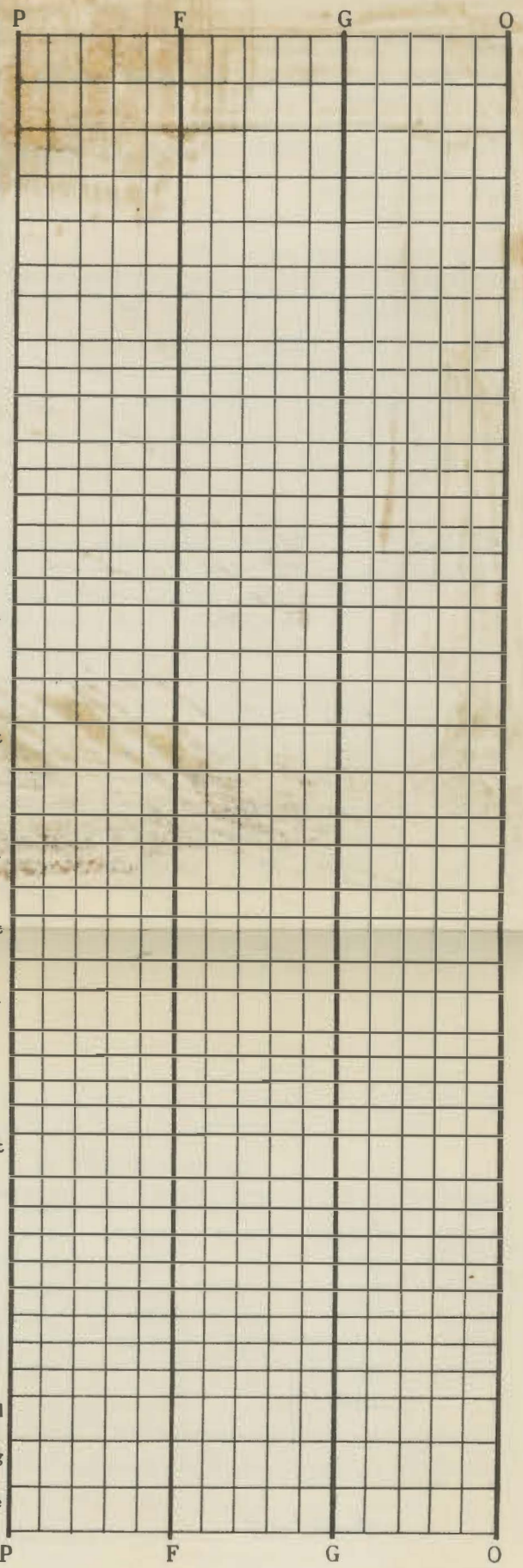
20. Avoids talking down to students.
21. Impartial in his treatment of students.
22. Gives well organized lectures.
23. Makes sure students understand difficult points.
24. Tolerant.
25. Gives regular quizzes.
26. Inspires you to make the maximum preparation for each day's assignment in his class.
27. Makes major points clear.
28. Willing to help students with research.
29. Never appears annoyed when approached outside of class.
30. Conscientious and thorough.
31. Makes his quizzes and exams agree with course objectives and content.
32. Interrupts a student answering a question only when necessary to clarify or summarize.
33. Avoids poking fun at people.
34. Paces lectures properly in speed and content to students' comprehension.
35. Conducts classes in an assured, confident manner.
36. Avoids talking to the wall or blackboard.
37. Usually relaxed rather than tense.
38. Enthusiastic in his teaching.
39. Unifies the subject in his lectures.
40. Avoids heckling students with long and close questioning.
41. Avoids being "too strict."
42. Analyzes problems clearly.
43. Holds the interest of more than just the brightest pupils.
44. Avoids being stern or dominating.
45. Seldom shows "It's right or completely wrong" attitude.
46. Controls class discussions by interpreting what has been said.
47. Uses vivid supporting details in lectures.
48. Impresses people favorably on first acquaintance.
49. Expects suggestions; but knows it is his business to direct activities.
50. Directs discussion skillfully.
51. Ends lectures effectively.
52. Maintains helpful attitude even when student is unprepared.
53. Avoids embarrassing or ridiculing a student in the presence of others.
54. Uses relevant supporting details.
55. Presents various aspects of controversial issues and encourages students to form own opinions.
56. Good physique.
57. Willing to admit errors in grading and change grades on the records.
58. Interprets abstract ideas and theories clearly.
59. Lectures well without depending on notes.
60. Avoids erasing useful material from the blackboard too soon.
61. Associates with students without becoming too familiar or intimate.



APPENDIX G

VALIDATION SCALE (PAGE 3)

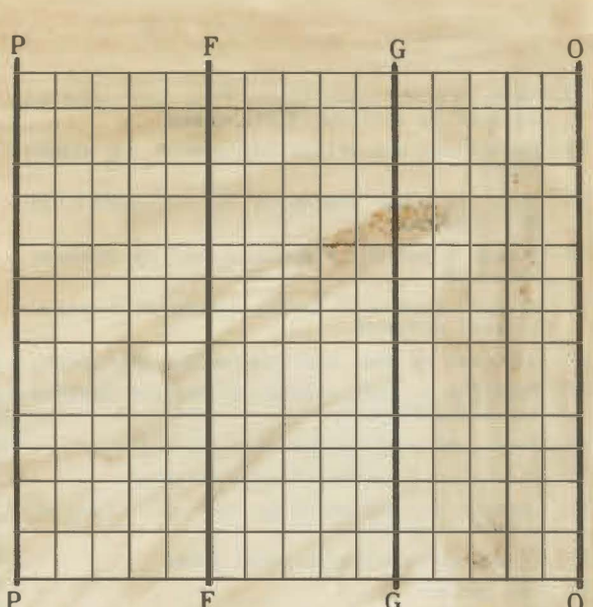
62. Makes clear transitions from one point to another by showing relationship.
63. Emphasizes and writes key points on blackboard.
64. Sets the same standards of attainment for all students.
65. Avoids appearing inconvenienced by students' requests.
66. Guides class well in discussion of controversial subjects.
67. Lectures so that students can take good notes.
68. Provides a good proportion between lecture and discussion.
69. Dispenses with unnecessary formality.
70. Sticks to a job until it's finished.
71. Teaches so that students' out-of-class interest is aroused.
72. Avoids overworking certain ideas.
73. Courteous.
74. Uses effective figures of speech.
75. Has well organized plans for the course.
76. Conducts classes informally.
77. Punctual for appointments.
78. Takes charge and leads class without lost motion.
79. Encourages initiative on part of students.
80. Contributes his free time generously to student activities.
81. Controls class discussions so as to prevent individuals from being hurt.
82. Presents a subject so that it is fresh and vital.
83. Uses apt and concrete examples of principles.
84. Permits making up quizzes or exams when absence is explained.
85. Has calm, even temperament.
86. ~~Answers student questions directly without~~ evasion.
87. Does not waste time or words.
88. Uses class average as standard instead of arbitrary standard.
89. Avoids using threat of poor grades.
90. Meets definite objectives of course.
91. Explains difficult material clearly.
92. Has easy yet forceful personality.
93. Addresses students as "Mr.--" or "Miss" so that they like it.
94. Emphasizes significant points with his voice.
95. Has ability to think "on the spot."
96. Keeps temper.
97. Praises more often than he criticises.
98. Makes assignments challenging and valuable.
99. Able to impart his enthusiasm to students.
100. Avoids trying to be impressive.
101. Refrains from talking about himself.
102. Has spent considerable time in research and study for each lecture.
103. Controls class discussions to prevent rambling and confusion.
104. Gives enough time on exams so that the average student can complete them.



APPENDIX G

VALIDATION SCALE (PAGE 4)

105. Avoids being "bossy."
106. Makes laboratory work meaningful rather than routine exercises.
107. Gives concise instructions.
108. Explains his position but does not argue with students.
109. Avoids shouting.
110. Has effective vocabulary.
111. Prepares summary carefully.
112. Removes excessive emotional pressure on student in discussion by shifting to another student.
113. Improvises training aids when necessary.
114. Analyzes class difficulties thoroughly.
115. Has many novel and stimulating ideas and viewpoints.
116. Aware that other people desire recognition and credit.



Further Instructions: Read *all* the phrases in the scale below and *then* check (✓) the *one* box best expressing your opinion of this teacher *as a person.* Please be completely frank. Indicate how you feel about this person rather than how you think you *should* feel. *Do not* turn back to read the Check List before you rate. Rate him as a *person; not* as a *teacher.*

AS A PERSON

FINEST PERSON I EXPECT TO KNOW-- WISH I WERE JUST LIKE HIM	*VERY FINE PERSON* --WISH MORE PEOPLE WERE LIKE HIM	*FINE PERSON* --WISH MORE PEOPLE HAD SOME OF HIS QUALITIES	*LIKED HIM VERY MUCH--* BETTER THAN MOST IN MANY WAYS	*ABOUT LIKE MOST PEOPLE* --WE GOT ALONG ALL RIGHT	*HE HAD SOME UNFORTUNATE QUALITIES--* HE ANNOYED ME OCCASIONALLY	*UNFORTUNATE PERSON* --DISLIKED HIM SOMETIMES	*UNPLEASANT PERSON* --I DISLIKED HIM	*EXTREMELY UNPLEASANT--* WORST I EVER KNEW OR EXPECT TO KNOW

Further Instructions: Read *all* the phrases in the scale below and *then* check (✓) the *one* box best expressing your judgment of his (her) *effectiveness as a college teacher. Do not* turn back to read the Check List before you rate.

EFFECTIVENESS AS A COLLEGE TEACHER

THE WORST TEACHER I HAVE EVER KNOWN	*DEFINITELY NOT WORTH* HIS SALARY AND MY TIME	*HARDLY WORTH* HIS SALARY AND MY TIME	*JUST WORTH* HIS SALARY AND MY TIME	*DEFINITELY WORTH* HIS SALARY AND MY TIME	*MORE THAN WORTH* HIS SALARY AND MY TIME	*A SUPERIOR* TEACHER	*ONE OF THE TWO OR THREE* BEST TEACHERS I HAVE KNOWN	*THE VERY BEST* TEACHER I HAVE EVER KNOWN

Thank you! _____

APPENDIX H

M A N U A L O F D I R E C T I O N S

for the

Rating Scale for College Teachers

University Instructor Rating Research

Sponsored by the Department of Psychology

University of Tennessee

Please direct all correspondence to:

E. B. Cobb
C/o Provost Marshal General's School
Camp Gordon, Georgia

MANUAL OF DIRECTIONS

1. Background and Purpose of the Rating Scale. The "Rating Scale for College teachers" began as a PhD thesis at the University of Tennessee. A large number of words and phrases descriptive of good and bad (by student definition) college teaching were collected by a variety of methods. The 1,000 plus items so collected were collated and reduced to 394 items by use of judges. The items were then formed in six Qualification Check Lists. A copy of one of these lists is attached as an exhibit. During the winter and spring of 1952 the check lists were administered in 124 classroom in 70 different colleges and universities scattered throughout the United States. Correlations (product-moment) were then calculated for: each item and the criterion rating of teacher personality; each item and the criterion rating of teacher effectiveness. The rating scale was then constructed by pairing items on the basis of their power to discriminate teacher effectiveness. Items low in discriminating power were paired with moderate ones, and moderate items were paired with these high in discriminating power. The lowest correlation used was 0.23. These paired items were selected so as to maintain approximate equal characteristicness (mean value on the 5-point check list, VG-VP scale). Research needs prevent showing this pairing in the scale. The research purposes underlying the administration of the scale are: cross validation of the item correlations, the collection of norms, and the testing of certain principles of rating scale methodology. Readers familiar with "forced choice" theory will recognize certain of these principles. A report of results

will be furnished participants in the research.

2. Use of the Scale. It will be recognized that the scale was built primarily for purposes of applied research. However, it should also be recognized that the scale is an effective instrument as it now stands. It may be used for administrative purposes or it may be used to educate teachers on the state of student opinion in their classes. No attempt is made here to define good rating practice. From the research point of view, the author is interested in various uses. Any local arrangements for use of the scale satisfactory to the user are satisfactory to the author, except as noted below.

3. Restrictions on the Use of the Scale.

(1) Student raters must NOT be told that "the scale is administered for research purposes." (See "Instructions to Students" below.)

(2) It is necessary to obtain ratings of the relatively poor teacher. If administration of the scale is done solely on individual teacher option, results will be difficult to interpret. Therefore, the smallest unit that can be supplied free of charge is that of the department. (See Costs.)

(3) Students should not sign their names.

(4) Reproduction rights are reserved.

4. Instructions to Students. The way the scale is used will determine what is said to the student. Each department or larger administrative unit should prepare clear and straightforward directions for their own use in administering the scales. Students should be told,

specifically, why they are to rate: (1) for the teacher's information, (2) for administrative purposes, (3) for both purposes. Please provide the author with a verbatim copy of the instructions used. (See Administration Log.) Again, do not say, "the purpose is research."

Do not inform students that the items are paired for research purposes.

The instructions must be brief if the scale is to be administered in 50 minutes.

5. Details of Administration.

(1) Each teacher in the department should be rated once by a class he is currently teaching. Use of the first class taught after 7:00 A.M. Monday morning is urged in order to control possible bias.

(2) "course Title," in the scale heading may be omitted by the student, since this entry may violate teacher anonymity. However, please show the course title on the Administration Log which accompanies the returned scales.

(3) Administer all scales within a department in the smallest possible time interval.

(4) Before administration, the last line of the scale (opposite Thank You!) should be completed with appropriate name or title -- head of department, for example.

(5) Do not permit unsupervised completion of the scale by students outside of class. An observer, not necessarily the teacher, should be able to complete the Administration Log accurately.

(6) Administration time is 50 minutes.

(7) Provide pencils for students. Page 4

6. Scoring and Interpretation of Results. Two kinds of scoring will be discussed: author scoring and local scoring. The author will score, analyze results, and provide each participating department with an interpretation for each teacher rated. There is no charge for this service. The research report of these results will preserve the anonymity of the colleges, departments, and teachers concerned. Naturally, such scoring, analysis, interpretation, and reporting must be delayed until all returns are in. It will take a few months to process the several hundred teachers needed.

Local scoring consists of tallying the results on a blank scale, computing mid-scores, and drawing a profile of mid-scores. It is suggested that the teacher rated be given the opportunity to do this. However, use of teacher code number and omission of course title makes it possible to do the clerical work administratively. Such scoring will aid the author materially and will shorten the processing time. Directions for local scoring are found on Pages 5 and 6.

7. Costs. Shipping costs will be paid by the author, both ways, in all cases. For participating departments, there is no charge for the scales. For individual teachers, the total cost, including scoring and interpretation, is 5 cents per student.

8. Shipping. Return all used scales by first-class mail. Four pounds is the weight limit for one package; use additional first-class packages for larger shipments. Do not insure. Weigh packages at post office and mark postage in ink. You will be reimbursed in 3-cent stamps

(checks for amounts over $1.00) unless directed otherwise. Please return unused scales. If there are 10 or more unused scales, please ship them by fourth-class mail.

9. Ordering. When requesting the rating scales, please list the number of classes and the enrollment in each. If additional copies of the Manual of Directions are needed, please indicate how many.

10. Directions for Local Scoring.

(a) The purpose in scoring is to find the mid-score for each numbered item on the scale. The mid-score is defined as N + 1 ÷ 2, where N equals the number of students in the class. If there are any omissions for a given item, it will be necessary to use the actual number of ratings made for that item.

(b) The scoring is done most quickly and accurately if one person reads the ratings from a completed scale and another person tallies these ratings on a blank scale. The reader should cut out the top row of boxes on a blank scale and place it over each completed scale. Number the boxes from 1 to 15, beginning with P. The reader then reads the number of the box checked by the student for Items 1 through 19.

(c) The recorder will find that an excellent method of tallying is to place the blank scale over a piece of cardboard; then use a pin to punch holes in the boxes as read by number by the reader. The punched holes are easy to count and take up little space.

(d) It will be easiest to read and record all the scales for Items 1 through 19, then repeat for Items 20 through 61, etc.

(e) Use the same method to tally the As a Person and Effectiveness As a College Teacher scales.

(f) T find the mid-score, count from either end of a row until you hit N + 1 ÷ 2. Disregard fractions. Your counting is made easy if you group your pin holes by five's, like this - ⁘ ⁘. Put a big pencil X through the box containing the mid-score.

(g) Pick up another blank scale and draw a profile of mid-scores on it. The tallied scale, together with the rating scales themselves, are required by the author for additional analysis, so please return them as shown under Shipping.

(h) If other statistics, such as the median or mean, are obtained, the author would appreciate a copy of the results, including the sum of each item row. It is emphasized that the completed scales (and tally sheets, if any) are the necessary items in the author's research, regardless of the methods employed in local scoring.

APPENDIX I

REPORT OF SCALE RESULTS TO DEPARTMENTS

(Letter of transmittal)

The Provost Marshal General's School
Camp Gordon, Georgia

June 24, 1954

Enclosed you will find sealed envelopes containing reports of the results of your participation in the University Instructor Rating Research project.

The project is far from complete, a fact which has made it necessary to avoid discussion of practical significance in the enclosed letter to participating teachers.

I hope you may find time to give me your suggestions on the administrative significance of the scale. The final scale will be much shorter and easier to score.

May I again express my appreciation for your interest and support.

Sincerely yours,

(signed)

E. B. Cobb

(Cover Letter to Instructors)

The Provost Marshal General's School
Camp Gordon, Georgia

June 25, 1954

Dear Colleague:

The enclosed material is an interim report to you, one of the college teachers participating in the research on the Rating Scale For College Teachers. The study is not yet completed; when it is, the published report will be called to your attention.

The first item is a scale containing your mean score for each item -- the right hand column of figures.

The next item is a mimeographed list of means for all teachers participating -- 71 in number, spread over 9 different schools.

I had hoped to be able to do the clerical work next involved for you, but I cannot meet the end-of-term deadline.

Using the "P" line as zero, and the first line to the right as 1, graph your scores on the scale. Next graph on the same scale the mean of the means from the mimeographed list.

You may thus see how much you deviate from the mean of each item. Regarding statistical significance: the standard deviations of the items vary, but the maximum standard error of the mean for the items was .238. Using a 1 per cent level of confidence, the probability is 99 to 1 that any deviation from the mean of .6 of a score point (6/10 of one cell on the scale) is not due to chance alone. Regarding practical significance: only you are in a position to make this interpretation.

As a matter of fact, I am deeply interested in your opinion on this question and would appreciate receiving your comments on the usefulness and validity of the scale. Please write as frankly and at as much length as you can.

Literally hundreds of thousands of tallies and machine operations have been made at this point, all done at personal cost. Your cooperation and interest has made the work possible. I intend to carry on

(Cover Letter to Instructors, continued)

the work and make a more complete report; your letters will provide encouragement and may help convince others that the research is worthy of financial aid.

Sincerely yours,

(signed)

E. B. Cobb

The Provost Marshal General's School
Camp Gordon, Georgia
E. B. Cobb

UNIVERSITY INSTRUCTOR RATING RESEARCH
ITEM MEANS -- TOTAL SAMPLE

Item No.	Mean of Means	Item No.	Mean of Means	Item No.	Mean of Means
1	10.08	39	10.20	77	11.10
2	10.03	40	11.23	78	10.66
3	10.00	41	11.01	79	10.07
4	10.84	42	10.27	80	9.94
5	11.05	43	10.36	81	10.46
6	10.31	44	11.17	82	9.92
7	10.49	45	10.85	83	10.54
8	11.40	46	10.50	84	10.67
9	11.27	47	10.14	85	11.53
10	10.41	48	10.73	86	10.81
11	10.35	49	10.44	87	10.10
12	10.32	50	9.82	88	10.01
13	10.66	51	9.39	89	11.21
14	9.81	52	10.25	90	10.82
15	10.56	53	11.11	91	10.17
16	11.44	54	10.50	92	10.68
17	11.37	55	10.33	93	10.51
18	10.69	56	9.77	94	10.19
19	10.34	57	10.36	95	10.93
20	10.70	58	9.88	96	11.69
21	10.75	59	10.50	97	10.43
22	9.87	60	10.93	98	9.45
23	10.12	61	10.97	99	9.50
24	10.80	62	9.80	100	10.58
25	9.62	63	10.23	101	10.72
26	8.72	64	10.71	102	10.47
27	10.41	65	10.96	103	10.39
28	10.29	66	10.19	104	10.22
29	11.40	67	9.38	105	10.75
30	10.92	68	9.49	106	9.98
31	10.70	69	11.06	107	10.30
32	10.83	70	10.62	108	10.81
33	11.28	71	9.57	109	12.13
34	10.17	72	10.08	110	11.52
35	11.06	73	11.68	111	10.14

UNIVERSITY INSTRUCTOR RATING RESEARCH
ITEM MEANS -- TOTAL SAMPLE (CONTINUED)

Item No.	Mean of Means	Item No.	Mean of Means	Item No.	Mean of Means
36	11.46	74	10.75	112	10.62
37	11.54	75	10.79	113	9.88
38	10.90	76	11.14	114	9.85
				115	10.00
				116	10.76

Data on "Person" and "Effectiveness" Nine-Point Scales

Person

Range: 7.2 to 3.4
Mean: 5.77
Standard Deviation: 1.2
Standard Error Mean: .14
1 per cent Confidence: .4 of Score Point

Effectiveness

Range: 7.4 to 3.1
Mean: 5.32
Standard Deviation: .94
Standard Error Mean: .11
1 per cent Confidence: .3 of Score Point

Distribution

"Person" nine-point scale distribution was rather skewed -- piled up toward the high scores. "Effectiveness" nine-point scale was approximately normal in shape.

APPENDIX J

SUMMARY OF VALIDATION DATA, QUALIFICATION CHECK LIST AND VALIDATION SCALE

Item No.	Item Person QCL	Item Person Scale	Item Person Dif.	Item Effectiveness QCL	Item Effectiveness Scale	Item Effectiveness Dif.	Couplet Difference Person QCL	Couplet Difference Person Scale	Couplet Difference Effective. QCL	Couplet Difference Effective. Scale	QCL-Scale Couplet Difference
1	.38	.29	.09	.28	.33	.05	.07	.16	.41	.25	.16
(2)	.45	.45	0	.69	.58	.11					
(3)	.71	.58	.13	.89	.76	.13	.07	.07	.26	.36	.10
4	.78	.51	.27	.63	.40	.23					
5	.19	.19	0	.27	.24	.03	.15	.23	.20	.21	.01
(6)	.34	.42	.08	.47	.45	.02					
(7)	.38	.21	.17	.55	.23	.32	.04	.02	.22	.06	.16
8	.34	.23	.11	.33	.17	.16					
9	.29	.22	.07	.29	.21	.08	.10	0	.22	.04	.18
(10)	.39	.22	.17	.51	.17	.34					
(11)	.60	.16	.44	.74	.37	.37	0	.14	.20	.03	.17
12	.60	.30	.30	.54	.34	.20					
13	.69	.52	.17	.59	.45	.54	.06	.08	.20	.05	.15
(14)	.63	.44	.19	.79	.50	.29					
(15)	.52	.35	.17	.66	.48	.18	.06	0	.27	.21	.06
16	.46	.35	.11	.39	.27	.12					
17	.40	.28	.12	.37	.28	.09	.14	.01	.29	.12	.17
(18)	.54	.27	.17	.66	.40	.26					
(19)	.73	.51	.22	.89	.41	.48	.01	.12	.34	.09	.25
20	.74	.39	.35	.55	.32	.23					
21	.58	.30	.28	.46	.22	.24	.02	.01	.47	.38	.09
(22)	.60	.31	.29	.93	.60	.33					
(23)	.73	.41	.32	.85	.55	.30	.05	.03	.33	.29	.04
24	.78	.44	.34	.52	.26	.26					

SUMMARY OF VALIDATION DATA, QUALIFICATION CHECK LIST AND VALIDATION SCALE (CONTINUED)

Item No.	Item Person QCL	Scale	Dif.	Item Effectiveness QCL	Scale	Dif.	Couplet Difference Person QCL	Scale	Effective. QCL	Scale	QCL-Scale Couplet Difference
25	.44	.11	.33	.28	.06	.22	.07	.29	.50	.31	.19
(26)	.51	.40	.11	.78	.37	.41					
(27)	.65	.40	.25	.95	.56	.39	.01	.07	.44	.24	.20
28	.66	.33	.33	.51	.32	.19					
29	.66	.50	.16	.46	.39	.07	.04	.02	.32	.11	.21
(30)	.62	.52	.10	.78	.50	.28					
(31)	.59	.19	.40	.78	.32	.46	.03	.13	.28	.11	.17
32	.56	.32	.24	.50	.21	.29					
33	.51	.18	.33	.40	.03	.37	.12	.17	.45	.39	.06
(34)	.63	.35	.28	.85	.42	.43					
(35)	.46	.38	.08	.62	.42	.20	.07	.16	.30	.29	.01
36	.39	.22	.17	.32	.13	.19					
37	.52	.33	.19	.48	.30	.18	.02	.22	.31	.25	.06
(38)	.54	.55	.01	.79	.55	.24					
(39)	.65	.31	.34	.74	.37	.37	.09	.04	.32	.25	.07
40	.56	.27	.29	.42	.12	.30					
41	.59	.24	.35	.46	.12	.34	.03	.11	.47	.39	.08
(42)	.62	.35	.27	.93	.51	.42					
(43)	.66	.65	.01	.85	.65	.20	.02	.26	.38	.41	.03
44	.68	.39	.29	.47	.24	.23					
45	.50	.30	.20	.41	.29	.12	.02	.05	.25	.08	.17
(46)	.52	.35	.17	.66	.37	.29					
(47)	.48	.58	.10	.93	.58	.35	.08	.09	.46	.35	.11
48	.56	.49	.07	.47	.23	.24					
49	.56	.32	.24	.54	.21	.23	.10	.14	.33	.23	.10
(50)	.66	.46	.20	.87	.44	.43					

SUMMARY OF VALIDATION DATA, QUALIFICATION CHECK LIST AND VALIDATION SCALE (CONTINUED)

Item No.	Item Person QCL	Item Person Scale	Item Person Dif.	Item Effectiveness QCL	Item Effectiveness Scale	Item Effectiveness Dif.	Couplet Difference Person QCL	Couplet Difference Person Scale	Couplet Difference Effective. QCL	Couplet Difference Effective. Scale	QCL-Scale Couplet Difference
(51)	.55	.40	.15	.73	.44	.29	.10	.03	.27	.27	0
52	.65	.37	.28	.46	.17	.29					
53	.58	.28	.30	.35	.14	.21	.10	.22	.36	.20	.16
(54)	.48	.50	.02	.71	.34	.37					
(55)	.35	.22	.13	.62	.14	.48	.02	.19	.33	.15	.18
56	.37	.41	.04	.29	.29	0					
57	.62	.22	.40	.58	.22	.36	.07	.12	.31	.37	.06
(58)	.69	.34	.35	.89	.59	.30					
(59)	.34	.28	.06	.60	.46	.14	.03	.06	.20	.20	0
60	.37	.22	.15	.40	.26	.14					
61	.62	.39	.23	.51	.30	.21	.04	.01	.22	.17	.05
(62)	.58	.40	.18	.73	.47	.26					
(63)	.35	.15	.20	.50	.30	.20	.07	.02	.22	.15	.07
64	.28	.17	.11	.28	.15	.13					
65	.68	.35	.33	.59	.24	.35	.06	.07	.20	.20	0
(66)	.74	.42	.32	.79	.44	.35					
(67)	.50	.14	.36	.76	.40	.36	.01	.21	.20	.01	.19
68	.51	.35	.16	.56	.39	.17					
69	.44	.26	.18	.34	.22	.12	.01	.24	.26	.13	.13
(70)	.45	.50	.05	.60	.35	.25					
(71)	.65	.52	.13	.78	.54	.24	.03	0	.27	.16	.11
72	.62	.52	.10	.51	.38	.13					
73	.71	.39	.32	.45	.19	.26	.02	.03	.29	.20	.09
(74)	.69	.42	.27	.74	.39	.35					

SUMMARY OF VALIDATION DATA, QUALIFICATION CHECK LIST AND VALIDATION SCALE (CONTINUED)

Item No.	Item Person QCL	Item Person Scale	Item Person Dif.	Item Effectiveness QCL	Item Effectiveness Scale	Item Effectiveness Dif.	Couplet Difference Person QCL	Couplet Difference Person Scale	Couplet Difference Effective. QCL	Couplet Difference Effective. Scale	QCL-Scale Couplet Difference
(75)	.51	.37	.14	.71	.50	.21	.10	.06	.34	.29	.05
76	.41	.31	.10	.37	.21	.16					
77	.20	.17	.03	.24	.16	.08	.24	.09	.47	.17	.30
(78)	.44	.26	.18	.71	.33	.38					
(79)	.62	.38	.24	.78	.34	.44	.02	.03	.33	.11	.22
80	.60	.35	.25	.45	.23	.32					
81	.65	.38	.27	.47	.20	.27	.09	.16	.48	.40	.08
(82)	.74	.54	.20	.95	.60	.35					
(83)	.59	.47	.12	.79	.54	.25	.13	.24	.41	.21	.20
84	.46	.23	.23	.38	.23	.15					
85	.69	.48	.21	.50	.18	.32	.01	.03	.33	.17	.16
(86)	.68	.51	.17	.83	.35	.48					
(87)	.46	.35	.11	.60	.50	.10	.06	.05	.28	.20	.08
88	.40	.30	.10	.32	.30	.02					
89	.63	.40	.23	.47	.33	.14	.05	.08	.44	.14	.30
(90)	.58	.52	.06	.91	.47	.44					
(91)	.93	.38	.55	.93	.44	.49	.12	.14	.30	.06	.36
92	.81	.52	.29	.63	.50	.13					
93	.37	.29	.10	.28	.08	.20	.09	.12	.31	.33	.02
(94)	.46	.45	.01	.59	.41	.18					
(95)	.48	.56	.08	.69	.65	.04	.04	.17	.35	.50	.15
96	.52	.39	.13	.34	.15	.19					
97	.50	.55	.05	.42	.20	.22	0	.15	.43	.11	.32
(98)	.50	.40	.10	.85	.31	.54					
(99)	.69	.65	.04	.97	.59	.36	0	.24	.46	.38	.08
100	.69	.41	.28	.51	.21	.30					

SUMMARY OF VALIDATION DATA, QUALIFICATION CHECK LIST AND VALIDATION SCALE (CONTINUED)

Item No.	Item Person QCL	Item Person Scale	Item Person Dif.	Item Effectiveness QCL	Item Effectiveness Scale	Item Effectiveness Dif.	Couplet Difference Person QCL	Couplet Difference Person Scale	Couplet Difference Effective. QCL	Couplet Difference Effective. Scale	QCL-Scale Couplet Difference
101	.32	.27	.05	.35	.05	.30	.12	.08	.38	.32	.06
(102)	.44	.35	.09	.73	.37	.34					
(103)	.46	.37	.09	.58	.55	.03	.15	.23	.25	.41	.16
104	.31	.14	.17	.33	.14	.19					
105	.73	.46	.27	.51	.20	.31	.07	.16	.44	.24	.20
(106)	.66	.30	.36	.95	.44	.55					
(107)	.62	.32	.30	.79	.39	.40	.08	.15	.34	.10	.24
108	.54	.17	.37	.45	.29	.16					
109	.33	.30	.03	.23	.07	.16	.09	.12	.29	.30	.01
(110)	.42	.42	0	.52	.37	.15					
(111)	.56	.23	.33	.78	.32	.46	.02	.08	.37	.01	.36
112	.58	.31	.27	.41	.31	.10					
113	.71	.19	.52	.44	.28	.16	0	.15	.34	.16	.18
(114)	.71	.34	.37	.78	.44	.34					
(115)	.66	.47	.19	.81	.44	.37	.05	.03	.26	.07	.19
116	.71	.44	.27	.55	.37	.18					