EVALUATION IN A MODULAR SELF-PACED INTRODUCTORY CHEMISTRY COURSE: IMPROVING THE TESTING AND MEASUREMENT OF ACHIEVEMENT AND ATTITUDE

By

Eugene Joseph Kales

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Departments of Chemistry and Administration and Higher Education

1978

ABSTRACT

EVALUATION IN A MODULAR SELF-PACED INTRODUCTORY CHEMISTRY COURSE: IMPROVING THE TESTING AND MEASUREMENT OF ACHIEVEMENT AND ATTITUDE

By

Eugene Joseph Kales

The first portion of this research concerns the investigation and improvement of the reliability of multiple-choice chemistry tests as used in a modular self-paced introductory course. This course requires hundreds of reliable yet easily graded fifteen-item exams per term. The multiple-choice format is readily scorable, but four- and five-choice questions are of doubtful reliability.

Test reliability was investigated as a function of the guessing opportunity when the number of answer choices varied from four to infinity (the short-answer item is treated as a multiple-choice item with an infinite number of choices). Test reliability was also investigated as a function of two types of item content: problems and nonproblems. Reliabilities in these studies are estimated by parallel forms correlation coefficients between two exams administered successively to the same students. Fluctuation in true scores is responsible for a drop of only 0.02 reliability units over the two hour testing inverval.

Eugene Joseph Kales

Guessing is a function of test difficulty -- students guess more often on a difficult test -- and when difficulty is taken into account, the correlation between number of answer choices and guessing error is over 0.90. Each doubling of the number of choices cuts guessing error approximately in half. Yet even when guessing is theoretically eliminated as in the short-answer item, reliability is not unity because parallel forms of an exam are not perfectly equivalent. This nonequivalence error is of the same magnitude as the guessing error for six-choice questions and it is relatively constant regardless of the number of answer choices.

At least in chemistry, whether test questions are problems or nonproblems also affects test reliability. Although both types of items show increasing reliability as the number of choices is increased, problems have higher average reliabilities than nonproblems. For the tests used here, problems averaged about twelve percent more reliable than did nonproblems. When the mix of problems and nonproblems is that which actually appears on exams used in these courses, reliabilities are midway between those of problems and nonproblems. Unreliability due to nonequivalence between parallel forms is also highest for nonproblems, lowest for problems, and intermediate for tests of mixed item types.

Increasing the number of answer choices from four and five to eight and ten produces sufficient improvement in test reliability.

Several examples of increasing the number of responses are discussed.

Two other aspects of the course which were investigated are the costs of the method of instruction and the student attitudes toward the course and subject.

The high initial and fixed costs of this system produce savings only when course enrollments are high. Costs are about twenty-nine dollars per student when enrollments are over fifteen hundred and increase to forty-three dollars per student as course enrollment declines to seven hundred. In comparison with a large lecture system, dollar savings are marginal, but instructional personnel are much more efficiently used.

Attitude survey results show a correlation near 0.70 with percent return and a correlation with course grade ranging from zero to 0.30. Observed response means are lower than typical university courses because students are not rating a person and because grades in the course average 0.24 units (on a four point scale) lower than student term grade point averages. A comparison of positively stated Likert items with 'naturally unbiased' evaluative items showed little difference between formats. What difference was observed tended to favor the Likert rather than the evaluative format.

ACKNOWLEDGEMENTS

I wish to express sincere thanks to Dr. Robert N. Hammer for his patient direction and constant encouragement through the long years of my graduate study. I also wish to dedicate a special thank you to the memory of Dr. William E. Sweetland. Without his initial support of and interest in my program, this joint degree would not have been possible. I wish to acknowledge the role of Dr. Robert L. Ebel in the molding of my beliefs and skills in measurement both as a member of my guidance committee and as an instructor in courses I have taken.

I wish to thank the Department of Chemistry at Michigan State University for its support in the early stages of my graduate study, and for the opportunity to meld research with teaching in an introductory chemistry course.

Finally, I wish to thank my parents, Frank and Genevieve Kales, for their unfailing support and loving acceptance.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter I Chemistry learning systems

I. A. Introduction

Teaching and chemistry are longstanding activities of man. Yet as sciences, both are relatively new. The science of chemistry is barely two hundred years old, and that of education is just out of infancy. Even newer is the application by chemists of instructional principles to their own teaching. As Havighurst [1] observed in 1941, "...chemistry teachers have not treated their teaching problem as a scientific one, deserving the same intelligence and industry, scientific method, and attitude that they devote to research in chemistry."

The years since World War II have been ones of burgeoning enrollments in college chemistry, coupled with modern budgetary limitations. Thus squeezed from both sides, the chemical educator is more concerned than ever in the past with the efficacy of his teaching methods. The learning system can be as simple as an instructor meeting daily with a group of thirty students, or as complex as a course with several lecturers, graduate assistants, proctors, graders, and a thousand students. In any case the teacher is concerned with the efficacy and efficiency of his system -- the results in student achievement, attitude, and behavior -- and the costs in time, money, and effort. Essential to any evaluation of a teaching method is the measurement of student attitudes and achievements. This study encompasses the analysis and refinement of the measurement techniques used in the recently implemented modular self-paced

introductory chemistry sequence at Michigan State University. Before a discussion of the experimental work is presented, a brief history of the methods of teaching chemistry and the recent developments in instructional technique are given, followed by a description of the instructional setting in which this work was done and to which these results apply.

I. A. 1. Historical methods of teaching chemistry

In days of yore the student learned his alchemy as an apprentice to a master chymist among the alembics and retorts of the master's laboratory. As the science of chemistry matured, the number of students to be instructed grew, and the tutorial master-apprentice method proved exceedingly inefficient. It was also inappropriate for the majority of students who did not aspire to be professional chemists, but who pursued chemistry for its cultural value or in the course of medical studies. The lecture became the principal mode of transmitting chemical knowledge.* Soon textbooks were being written as supplements and aids to the course lectures -- the first entirely domestic chemistry text was written in 1819 by Gorham.

During these early years, students were almost never provided with laboratory instruction. Benjamin Silliman [3], the most prominent educator of his day, reflecting his time said of students,

*For a complete treatment of the historical development of chemical education in America see references 2-5.

"...I should much prefer that they should do nothing; for then they would not hinder me and my trained assistants, nor derange nor break the apparatus."

Near the close of the nineteenth century, the value of laboratory instruction was more widely accepted. About this time the chemistry graduate student made his appearance, and by 1914 one-seventh of the Ph.D.'s granted by American universities were in chemistry. Coupled with the greater availability of laboratory apparatus, these factors led to rapid and nearly universal inclusion of chemistry laboratory instruction as part of the learning system. Of somewhat slower spread was the recitation or quiz section, but by the nineteen-twenties it too became a standard piece of educational machinery.

A survey taken by Hendricks [6] in 1924 found that chemistry students were afforded two or three lecture hours per week, one quiz hour, and three or four laboratory hours. The quiz or discussion sections along with the laboratory meetings have not varied much in size between fifteen and thirty students from the earliest days of chemical instruction to the present. In contrast, lecture enrollments have increased until four hundred in one lecture arena is not an unknown occurrence. As early as 1930 Ehret [7] described the beginning chemistry courses at New York University with annual enrollments of nearly 1400 students employing nine lecturers and thirty graduate assistants. Today most major institutions contend with similar numbers.

To accommodate large heterogeneous enrollments, educators in many disciplines have experimented with teaching methods sometimes radically different from the traditional lecture format. Most of these innovations awaited the development of the technology to release the instructor from live oral delivery or printed textual material. This technology began influencing college courses from the nineteen-fifties onward.

I. A. 2. Modern innovations in pedagogic technology

The postwar progress in the technologies of communication and calculation has had a major impact on how courses are taught. The major advances include television, high-speed computers, audio and video cassettes, optical scanners, teletypes and terminals, and low-cost duplicating machines. The three most important learning system innovations are (1) audio-visual-tutorial, (2) programmed instruction, and (3) self-paced instruction. Each of these innovations will be briefly discussed along with the technology which made them possible.

I. A. 2. a. Audio-visual-tutorial instruction

S. N. Postlethwaite [8, 9] introduced the audio-tutorial system at Purdue University in 1961 as a program of remedial assistance for introductory botany. The heart of the system is the prerecorded course material to which students have free access. Initially and still

most often this material is an audio cassette recording which serves as a flexible substitute for the lecture. Postlethwaite [10] soon expanded the method beyond simple remedial reviews to include complete instructional modules. Each audio-visual-tutorial module includes audio cassettes, visual aids, supplementary printed material, and experimental equipment or models.

Two outstanding benefits distinguish AVT methods from the traditional lecture. First, the student is not restricted in his study of the course material by the schedule of lectures. Students may listen essentially anytime (often including evenings and week-ends) to the audio cassettes. Second, as a student is going over the material, he is not constrained to cover it at the same speed as another student, but may stop, race ahead, or backtrack as he feels necessary. These two reasons may explain why television never approached the popularity of the audio cassette.

Television as a teaching aid was reviewed by Barnard [11-13] in 1968. Televised lectures provide none of the benefits of the audio cassette, and retain the drawbacks of the large lecture in addition to the impersonality of a recorded presentation. Video cassette equipment is still too expensive and cumbersome for widespread use; thus television is limited to areas where the visual image is integral to the presentation. The most successful continuing use of television has been in laboratory instruction. When used to demonstrate the procedures to be used and show expected results, a picture is worth

a thousand words. The televised pre-laboratory program maintains a consistency of presentation unmatched by the differences within a group of laboratory instructors. When finally video cassettes are as inexpensive and easy to use as audio cassettes are today, then television will gain acceptance as a convenient option to the book, tape, or lecture.

I.A.2.b. Programmed instruction

Large lectures and the various media methods excel in the efficient presentation of material; they do not require but only encourage the participation of the student in this process. Programmed instruction is a systematic plan of presenting course material which requires the active participation of the student as he proceeds through the lesson. Jesse H. Day [14] summarized the promises of programmed instruction:

> "[the student] cannot proceed from item 1 to item 2 until he has learned item 1. Thus there is a built-in guarantee that when he has finished the program he has learned all of it, not just the 90% needed for an A.
>
> He is continuously active. He cannot skim or star-gaze; the program waits patiently -- the student must make active effort. The student learns what he does; he is no passive sponge.
>
> He proceeds at his own pace; the bright go quickly, the not-so-bright more slowly.
>
> The well-written program is so logical, clear, and consecutive that the student rarely fails to make each step. It is not like a turgid textbook whose explanations need explanations.
>
> The student receives immediate confirmation of each right answer, or immediate correction if wrong. There is no waiting while papers are graded, the question forgotten,

> and interest past. The confirmation of correctness at each step is tremendously encouraging to the student and provides confidence for the next step."

There are two fundamentally different kinds of programs -- linear and intrinsic. Jay A. Young [15] describes the linear program as a series of statements interspersed with fill-in questions. Each paragraph with a question is a 'frame' and the steps between frames are small enough that the student will almost never miss a question. The material is presented in a logical (linear) progression from the initial to the concluding statement. Day [16] describes the intrinsic program as branching or multiple-choice rather than linear or fill-in. The student is presented with a question and asked to select his answer from a set of responses. Each response leads the student to a different branch of the program, explaining the mistake which led to the wrong answer, or reiterating the reasoning which gave the correct answer.

The 'teaching machines' mentioned by Day [17] and Skinner [18] are automated versions of programmed instruction. S. L. Pressey [19] may be credited with inventing the first true teaching machines in the nineteen-twenties. However, the educational world was not then ready for such an industrial revolution. Pressey [20] despaired at this intransigence: "The writer has found from bitter experience that one person alone can accomplish relatively little, and he is regretfully dropping further work on these problems." It was not until learning theory provided a rationale for the success of teaching machines

and other forms of programmed instruction that their use increased.

B. F. Skinner [21] discussed the postwar developments in the experimental analysis of behavior and concluded that the principles of conditioning, shaping, and behavior modification have an exciting future in education. "The principles emerging from this analysis, and from a study of verbal behavior based upon it, are already [1957] being applied in the design of mechanical devices to facilitate instruction in reading, spelling, and arithmetic in young children, and in routine teaching at the college level."

The most advanced form of the teaching machine is the computer. The different ways programmed instruction has been computerized are detailed by Castleberry [22], Ewig [23], Lower [24], and Grandey [25]. The use of computers requires a large initial investment in interactive computer terminals and the development of sophisticated computer programs. Perhaps because of this high cost in time and money, computerization has only had a supplementary role in instruction. Computer assisted instruction (CAI) is a very successful, if expensive, method of drill and practice. When the costs decrease and when successful packaged programs are readily available, the computer may play a larger part in instruction.

I. A. 2. c. Self-paced instruction

Fred S. Keller [26] described an innovative learning system in his seminal article of 1968. Keller summarized five distinguishing features which set his method apart from the conventional:

(1) The student goes through the course material at his own pace
(2) Students cannot proceed to new material until they have demonstrated mastery of previous material
(3) Lectures and demonstrations are supplementary and motivational, and do not contain any essential material
(4) The written word is stressed in teacher-student communication
(5) The use of assistants and student proctors permits repeated testing, immediate feedback, and almost unavoidable tutoring

The Keller plan, or Personalized System of Instruction (PSI) as it is now most often called, has spread with almost revolutionary fire. Indeed, a scant three years after Keller's original article, Green [27] asked whether the plan was catching on too fast. Many modifications have appeared, but the most enduring feature in these variations on the Keller theme is self-pacing.

The self-pacing element in PSI demands that an alternative to the traditional fixed class meeting be found. Keller points out that this does not mean the teacher must use programmed instruction, audio-visual media, or teaching machines. What is most important is that students be told exactly what they are expected to learn and then be given the freedom to determine when and how to learn it. The Keller plan does not depend on modern technology. Whetzel [28] described in 1930 a 'Keller-before-Keller' method relying on a textbook and mimeographed handouts. It seems again, as in the case

of programmed instruction, that growth of the self-pacing method waited until the more favorable climate of the last decades.

I.A.2.d. The confluence of innovations

Yesterday's teacher had few choices to make about the methods of his profession. Beyond deciding whether to give three or two hour exams, how much to weight the final, which demonstration to perform in lecture, and other essentially minor concerns, the instructor practiced within the time-tested traditional format. Today's teacher has many more opportunities to be the creative manager of a learning system rather than a parrot or entertainer or disciplinarian.

The options available to the instructor planning a course are many, and his approach to designing a learning system will be eclectic. A list of the major choices afforded the instructor is presented in Table 1.1 in four areas -- presentation, pacing, testing and grading, and assistance.

Table 1.1 Instructional design options

Presentation of course material

1. Large lecture - little or no teacher-student interaction
2. Small lecture - opportunity for questions or discussion
3. Seminar - discussion without formal lecture presentations
4. Audio cassettes available for unscheduled student use
5. Televised lectures live or on a limited schedule
6. Video cassettes available for unscheduled student use
7. Programmed texts for student self-study or remediation
8. Teaching machines for student self-study or remediation
9. CAI - interactive computer system of programmed instruction
10. Textbook - usually as an adjunct presentation
11. Handouts - as supplementary material

Pacing of presentation

1. Instructor pacing according to a definite schedule
2. Self-paced within the matriculated term
3. Modular pacing - self-pacing between scheduled deadlines

Testing and grading

1. Single-try examinations with a 'curved' grading scale
2. Repeatable mastery tests with passing score set very high - grade is based on the number of tests successfully passed
3. Repeatable tests graded with a predetermined scale - course grade is based on the cumulative or 'last-tries' average

Assistance

1. Interruption of class presentation
2. After class or by appointment
3. Special recitation or discussion sections
4. Help Room assistance
5. Supplementary presentation material
6. Tutorial assistance by contract or appointment

I.B. Introductory chemistry at Michigan State University

The curriculum of chemistry courses offered in the Department of Chemistry at Michigan State University underwent a major revision in the Fall of 1967. Previously the department provided two introductory sequences -- CEM 101, 102, and 103 for nonmajors and CEM 111, 112, and 113 for majors. Currently ten different courses are offered which can be arranged in more than ten different sequences. The hierarchy of paths among these courses is illustrated in Figure 1.1. Until 1973, however, all the courses were taught by the traditional lecture-recitation approach. In Winter of 1973, a major change was made in the way CEM 130 and 131 are taught. These two courses -- which enroll about one-third of the entire freshman class at the university -- are now taught by a modular self-paced system. Both the traditional methods and this recent innovation will now be described.

I.B.1. Traditional teaching methods

When the syllabus of introductory chemistry courses was restructured in 1967, the laboratory sections previously attached to specific lecture courses were established as separately numbered courses. The methods by which these courses were and continue to be taught are typical* and will not be further discussed.

*The only feature which might be considered in any way unusual is the occassional use of recorded television pre-laboratory demonstrations.

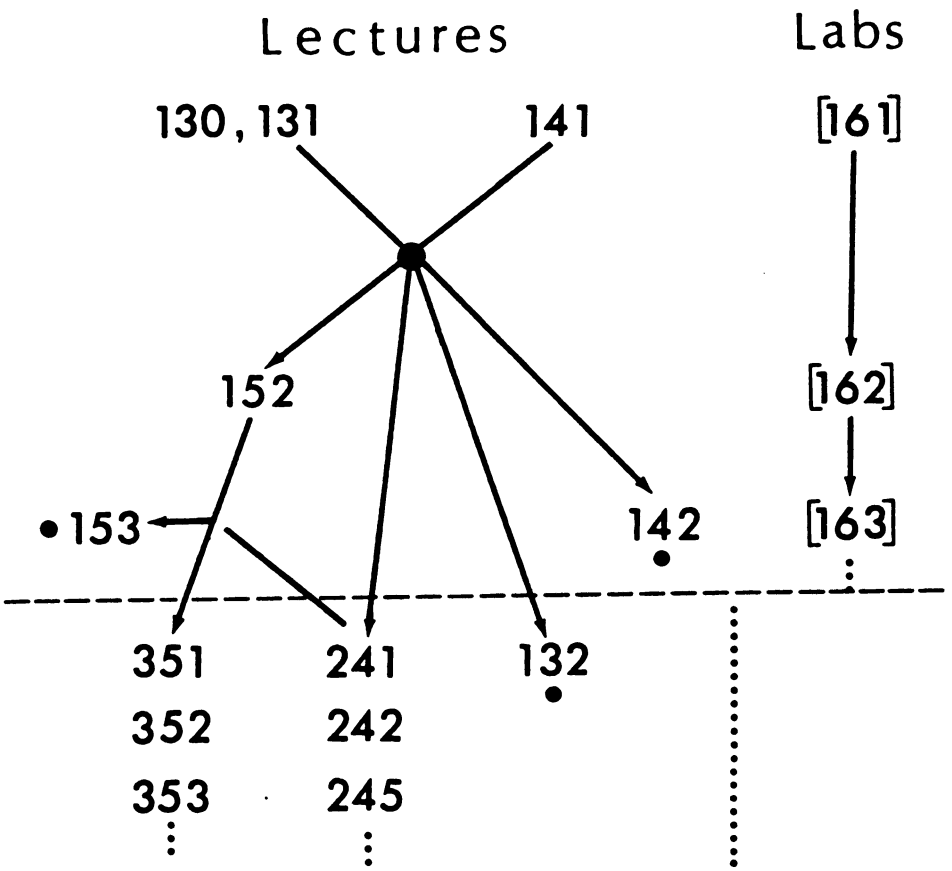


Figure 1.1 Hierarchy of introductory chemistry courses. Courses in the same row cannot both be taken for credit; unbroken arrows lead from prerequisite to successor courses. The dashed line separates general from organic chemistry; the dotted lines indicate paths of continuation in chemistry.

The description of the traditional learning system used in the lecture courses is organized according to the headings introduced in the previous section listing the choices afforded the instructor in the design of a learning system.

Presentation of course material. Traditionally, the live lecture is the basic vehicle used to present course material to students. When the size of the enrollment precludes all students meeting in the same room, separate-but-equal lecture sections are offered. The lecturers in multi-sectioned courses meet regularly and discuss what will be covered in the syllabus of topics in order that lectures proceed concurrently through the same material. Faculty members are assigned to lecture in the various courses on a rotating basis according to their subject competency and pedagogic preferences.

All the lecture courses use a textbook as a secondary method of presentation. The extent to which the lecturer relies on the text is a function of both the adequacy of commercially available textbooks and also of the instructor's style. Additional material is frequently presented as handouts reproduced in quantity in the Chemistry Department copy center. These handouts may be necessary to bridge gaps in textbook coverage of syllabus topics or may only be supplementary homework assignments and practice problems.

Pacing of presentation. Since the principal mode of presentation is the live lecture, all students are constrained to progress at the same rate. These lecture courses are almost entirely

instructor-paced; only when the instructor treats the lectures as secondary to the textbook does the student have any opportunity to alter the pace at which he covers the material.

Testing and grading. Measurement of student achievement is accomplished by lengthy and infrequent testing. Two or three hour exams and a two-hour final exam are used to obtain a distribution of student achievement. Courses which also have recitation or quiz sections may often include grades from homework or short quizzes in a student's total score.* Based on the distribution of total or mean scores, and also on the instructor's expectations of class performance, cutoff points are established and grades are assigned. The process ensures that, at least for large classes, the distribution of final grades is relatively constant from term to term.

Assistance. Most of the introductory chemistry courses have a large lecture section accompanied by a smaller recitation or quiz section. These recitation classes are staffed by graduate teaching assistants with whom about twenty-five students meet for one hour each week. During recitation, students have the opportunity to ask questions, discuss the lecture, and review homework problems. In addition, a general chemistry help room is staffed on a rotating basis by graduate assistants assigned to recitation sections in the largest

*Schwendeman [29] described a method of standardizing the grades received on quizzes or homework based on the common hour and final examinations. This computerized method of averaging and weighting was sometimes used in CEM 111 and 112 and in CEM 130, 131, and 141.

courses. Students may receive more individualized assistance on a walk-in basis by going to the help room. Also, special help sessions are usually held the day before important examinations. And finally, students may seek help from the lecturing professor or teaching assistant either after class, during office hours, or by appointment.

I.B.2. Modular self-paced introductory chemistry

In the spring of 1972 the Chemistry Department at Michigan State University accepted a proposal from the Educational Policies Committee and the General Chemistry Committee to completely restructure the ways in which CEM 130 and 131 were taught. As the principal architect of the new system, Dr. Robert N. Hammer [30] set out some of the philosophy and assumptions on which the planned changes were based. These are exerpted here:

> "The aim of this plan is to improve the teaching of introductory chemistry by maximum individualization of subject matter and teaching methods without burdening the faculty or graduate teaching assistants, without increasing costs, and without sacrificing scholarly standards."
>
> "Although there is nothing sacred about the way we teach our majors, neither is there any compelling need to change what we teach or how we teach it. Chemistry majors and others who are happy in the lecture-recitation system might be handled by continuation of the ... 141, 152, 153 sequence."
>
> "Lecturing is an inefficient way to teach facts and ideas to most beginning students."
>
> "Students should be told in detail and in writing exactly what they are expected to know and what they are to be able to do after studying a topic."

"Students can and will teach themselves when given guidance, proper learning resources, and adequate rewards. When each of these essentials is present, self-teaching appears to be more popular and more effective for most students (though probably not all) than is passive listening in the lecture room. Self-teaching has the additional advantage that it cultivates independence on the part of the student."

"There are never enough instructors and never will be. However, this proposal is based on the assumption that we probably must get along with fewer Graduate Teaching Assistants in the near future and that dollar costs can be no greater than now, once an operational steady state has been attained. The unmistakable message, then, to anyone who wants to improve chemical education is that people must be reserved for those tasks that need people, and people's skills and abilities must be focused on tasks they fit. Furthermore, machines -- of many kinds -- must be used wisely to do the repetitive, mechanical, or clerical tasks which we can no longer afford to do with professionally trained minds and hands."

"Another assumption of this proposal is that the 'equal time and treatment for all' philosophy characteristic of traditional classroom teaching is no longer acceptable."

"We should work toward an educational system in which a student can study in his individual way -- with guidance, learning aids, and rewards -- and receive individual help if and when he needs or wants it."

"The ability to work rapidly does not seem to have sufficient merit to compensate for damage done by rushing students through learning or testing processes. Neither does there appear to be much justification for the requirement that all students work at the same rate. Hence, within the broad limits imposed by the term system and economical program administration, students should be free to progress at self-chosen rates."

The system machinery for the two-term sequence of CEM 130 and 131 was constructed during the fall and winter of 1972-73, and teaching actually began during Winter and Spring of 1973. The basic design elements are discussed below; also noted are the evolutionary changes which were made during the period of these studies.

I.B.2.a. Syllabus and study guides

The syllabus of topics is the table of contents of these courses. It has been divided into units of work approximately equivalent to one lecture period plus associated out-of-class assignment. There are thirty-six units in CEM 130 and twenty-seven in CEM 131 so that nine units correspond to one academic credit. A study guide for each unit* describes the path by which the stated goals of the unit are reached. The study guide is the indispensible central element in this instructional system.

Each study guide consists of three parts. The first section is an introduction to the topic and a statement of the educational goals of the unit. Although early versions of some study guides included objectives set forth in behavioral terms, the current practice is to be more general. A more practical example of the expected student behavior is provided by the study questions and practice problems.

The second component of the study guide is titled What To Do. Four basic steps form the skeleton of this pedagogic algorithm. First, the student is to listen to an audio cassette for that unit. Second, he is to review the appropriate textbook references. Third, he should answer the study questions and work the practice problems included in the study guide. Finally, the student is directed to begin the next unit. This basic listen-read-practice-proceed sequence of steps is frequently

*Only in rare cases where subtopics cannot be divided into separate packages are more than one unit included in a single study guide. The syllabus is listed in Appendix A.

expanded or elaborated. The student's attention may be directed to special models or laboratory examples. A special motion picture or film loop may be recommended. Reminders are often made to review the syllabus and to seek assistance in the Help Room when needed. When a unit is the last in a series for an exam, the student is advised to try a sample exam, do any needed review, and then take the formal examination covering those units.

The list of study questions is the final element in each study guide. Occassionally, supplementary notes will be included with the study questions so that what is heard on tape or read in the text is reiterated in an easily reviewable form. These questions and problems provide a working example of the objectives of the unit. The sample exam included with the last unit in a testable cluster is yet another example of unit objectives. When the student is able to answer these questions he is ready to continue to the next unit or exam.

Thus the study guide fulfills three functions:

1. Define exactly what the student is to learn.
2. List the resources available to the student.
3. Provide means for self-estimation of competency.

The study guides continue to undergo many minor modifications as the courses are taught, but their basic structure remains unchanged. The syllabus was extensively rearranged during the 1976-77 academic year. Units in the two-term sequence CEM 130/131 are now numbered consecutively rather than beginning at '1' for the second course.

I.B.2.b. Presentation of material

The primary method of presentation in this system is the audio cassette tape. The instructor speaks to the student in a conversational tone and interspaces his presentation of the topic with examples and questions. Each tape is accompanied by a visual supplement -- called Tape Notes -- to which reference is made when a picture, table, graph, or written list is particularly helpful. The students are encouraged to stop the tape often, to backtrack if necessary, and to take notes freely. A special Tape Room currently houses over 150 cassette playback units -- up from an original thirty -- and the students may check out tapes and listen to them with headphones days and evenings during the week and on afternoons Saturday and Sunday. During peak hours, the only sounds heard in the Tape Room are the stops and starts of the audio machines and the whirs of rewind and fast-forward; it sounds much like a convention of crickets and click-beetles.

Students are not restricted solely to listening to tapes in the Tape Room. Beginning in Fall 1973 a tape duplicator was made available for general student use on a sidetable in the Tape Room. Within two years it was decided that operation of the duplicator would be restricted to trained personnel. Students who desired a copy of an audio tape would turn in a blank cassette in a self-addressed envelope and it would be transcribed and ready for pick-up the next day. This procedure soon became burdensomely cumbersome as the number of duplications per term neared twenty thousand. From Fall 1976 onward,

a student who wishes copies of tapes may get them from the Tape Exchange window. To receive a prerecorded cassette, the student must trade either special pre-paid scrip or a previous cassette obtained with such scrip.

A secondary resource in this system as in the lecture system is the textbook. The ordering of topics in the syllabus was prepared independently of any textbook; hence none exactly follows the progression of study units. When the new system was first implemented, only one textbook was used and the appropriate page and chapter references were included directly in the study guides for the various units. Two terms later, a second textbook was added to the roster and the students given their choice of which to use. Since it was forseen that textbooks might be changed often, references to textbook assignments were thereafter included on a separate sheet. In the past four years, five different textbooks -- at least two each term -- have been offered as options to the students.

The most significant characteristics of this means of presentation are its flexibility and consistency. Audio tapes are modified and improved with little more effort than needed for the preparation and delivery of a lecture. Once recorded, the audio cassette delivers the course material to any number of students with a consistent quality and a flexible rate and schedule. Additionally, the flexibility present in multiple textbook references allows the student to choose that book which most appeals to him. It also allows the instructor to change

textbooks easily since the sequence of topics has never been tied to the presentation of a single author.

I.B.2.c. Help Room assistance

The Help Room is intended to be a place where students can interact with instructors and with each other. It is open five days a week except Friday evening and is staffed with graduate teaching assistants according to the expected student attendance. Two quite different kinds of student-instructor activity occur -- depending on the number of students present.

When very few students are in the Help Room, the atmosphere is that of a quiet lounge. Casual discussions occur among students and instructors, and much real tutoring takes place. Students who want or need substantial individual assistance attend these off-hours regularly.

In contrast, when large numbers of students crowd into the Help Room, it seems a bedlam. Some students are still to be found quietly studying in twos and threes around the room. Each instructor is generally the focus of a knot of students seeking help (at times vociferously) or listening as if at a standing-room-only recitation class. Movement swirls through the room as students come and go, and as instructors move across the blackboard or search out those too timid to 'butt right in'. Two informal rules governing rush-hour instructors are (1) don't stand still, and (2) don't talk to only one student.

As a supplement to the Help Room, special help sessions are often scheduled to review material which is known to cause many students difficulty. These formal sessions provide a more traditional setting for the students best helped by a scheduled class.

I.B.2.d. Testing and grading

The examination system used in CEM 130/131 may be best described as a modified mastery system. Every certain number of units, the student is directed to take a proficiency exam. A predetermined grading scale is used to compute the grade on each exam. If the student is dissatisfied with his grade, he may retake the exam without penalty. Only the last try for each exam is the grade which counts toward the student's term average. Within the time limits when an exam is offered, the student is able to set his own level of mastery and retake examinations until he achieves this level. At the end of the course a cumulative final examination is given. The effect of the final exam on a student's grade is restricted to raising or lowering the grade one step.* Depending on what grade the student has when he takes the final exam, he must score some minimum on the final to keep the same grade or must score at or above some higher score to raise his grade. If the student scores within approximately ten percentage points of his term average his grade will be unchanged.

*The current university grading system is numerical between 0.0 and 4.0 in steps of 0.5 units with the exception that there is no grade of 0.5.

The grading scales used, the number of exams in a course, and the number of days each exam is offered have undergone considerable change in the more than four years the system has been operating. The initial design for these courses had nine exams in CEM 130 and seven in 131. The approximate 'windows' -- the days during which an exam is offered -- for each exam in the two courses are displayed in Figures 1.2 and 1.3. The original scheduling of examination windows has considerable overlap. There were many days when between the two courses seven different exams were being given.

The number of exams and the number of days each is offered were reduced to the current schedule as illustrated in Figure 3. Two benefits and one loss result from this adjusted arrangement. One gain is administrative. Previously, the multiplicity of forms and exams had often overloaded the exam-generation machinery. Now, fewer different forms need be generated of each exam, and fewer different exams are given and graded on any one day. The second gain affects students. The original scheduling of examinations allowed students to procrastinate to their detriment. Within the last ten days of classes, many students still needed to study for and take three separate exams. With fewer exams and a tighter schedule, it is much less likely that students will fall dangerously behind in their studies. There is, however, a major loss with the current schedule. The extent to which students are allowed to pace themselves is limited severely by the first and last days an exam is offered. The self-pacing in these

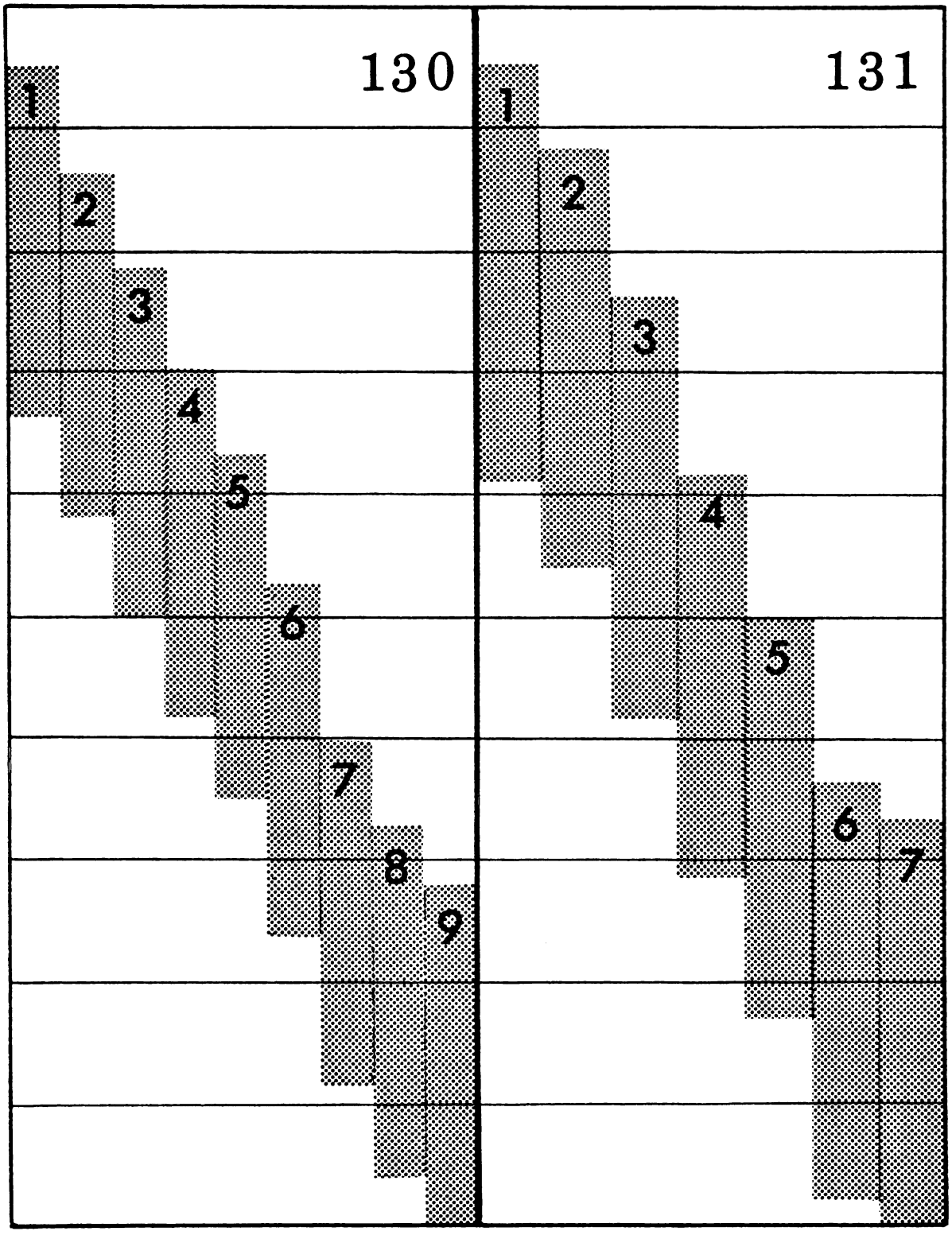


Figure 1.2 Original scheduling of examination windows. Each term block is divided into ten weeks.

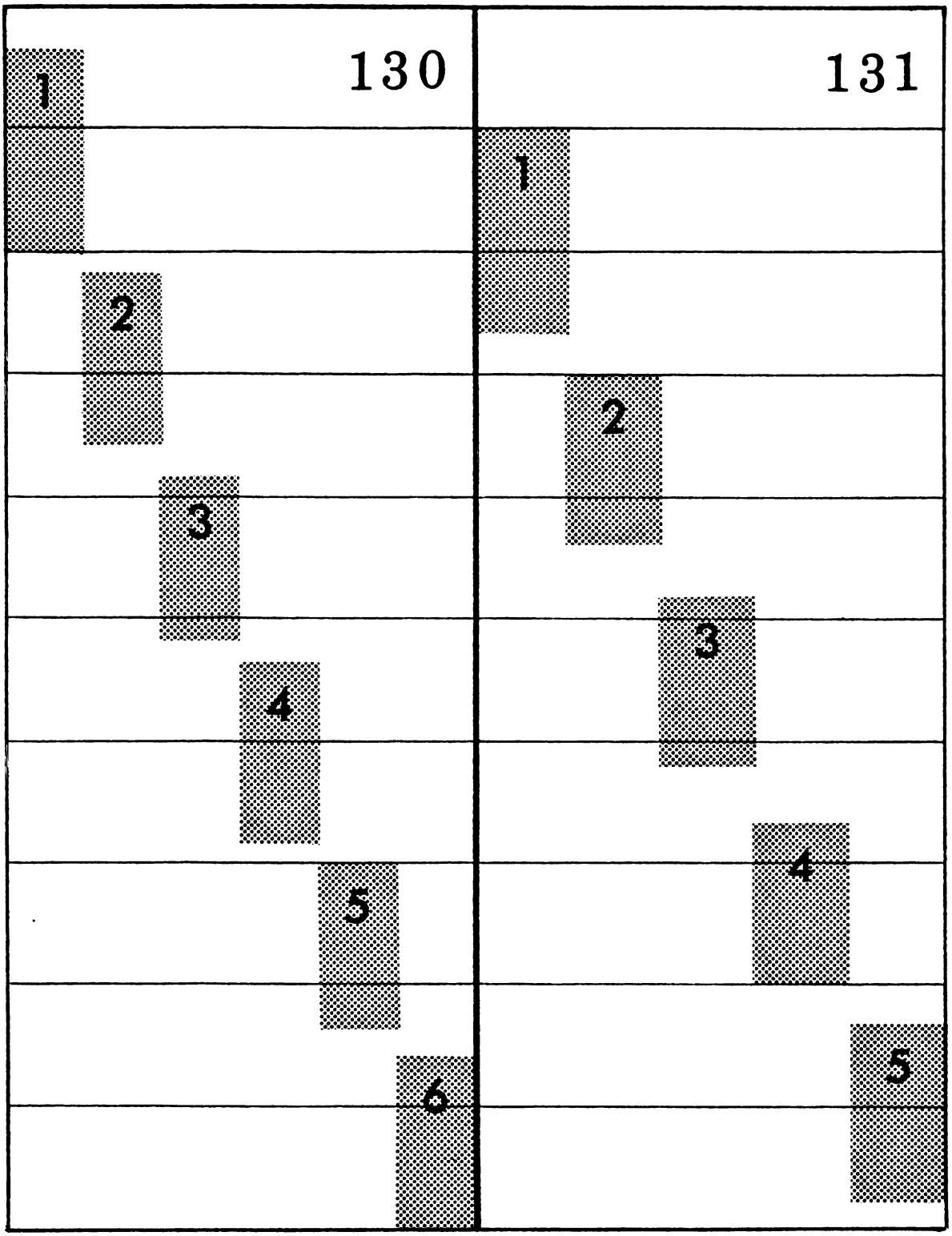


Figure 1.3 Current scheduling of examination windows. Each term block is divided into ten weeks.

courses is now more strictly modular than in the original design. Within each module defined by the closing date of an exam, the student studies and is tested at his own pace, but the modules themselves are now instructor-paced. It is planned for the near future, within the current framework of minimum-scheduled-progress, to allow students to proceed as rapidly as they are able by moving the opening dates for each exam nearer the beginning of the term.

The grading scales used in CEM 130/131 have also evolved from the original design to the present. The different scales used in these courses are displayed in Table 1.2. The initial cut-off points in the grading scale were derived from a composite grading scale for the previous four years. It was assumed that the tests to be given in the new courses would not be much more difficult or easy than they had been in the past. After the system was in operation for a year, a major upward shift in the scale was made. It was found that the abbreviated length and the no-fault repeatability of exams gave grade distributions with unjustifiably high means. After this major change in the scale, several minor adjustments were made during succeeding terms to accommodate mathematical peculiarities observed in the averaging of the fifteen-item examinations to give an overall percentage.

Table 1.2 Grading scales used in CEM 130 and 131

Terms	Grade						
	4.0	3.5	3.0	2.5	2.0	1.5	1.0
Minimum term average* needed to receive a particular grade							
130W73 through 131SS74	90.0	83.0	76.0	69.0	61.0	53.0	45.0
130F74 and 131F74	93.0	87.0	80.0	73.0	67.0	60.0	53.0
130W75 through 131SS75	92.0	86.0	79.0	72.0	66.0	59.0	52.0
130F75 and 131F75	92.0	85.0	79.0	72.0	65.0	59.0	52.0
130 from W76 onward	92.$\bar{2}$	85.$\bar{5}$	78.$\bar{8}$	72.$\bar{2}$	65.$\bar{5}$	58.$\bar{8}$	52.$\bar{2}$
131 from W76 onward	92.0	85.$\bar{3}$	78.$\bar{6}$	72.0	65.$\bar{3}$	58.$\bar{6}$	52.0
Minimum score on the final exam needed to raise a grade one step							
130W73 through 131SS74**	--	94.0	87.0	80.0	72.0	64.0	56.0
130F74 and 131F74	--	90.0	87.0	75.0	68.0	60.0	53.0
130W75 onward	--	94.0	90.0	82.0	75.0	67.0	60.0
Minimum score on the final exam needed to keep a particular grade							
130W73 through 131SS74**	80.0	73.0	66.0	59.0	51.0	43.0	35.0
130F74 and 131F74	75.0	68.0	60.0	53.0	45.0	40.0	37.0
130W75 onward	82.0	75.0	67.0	60.0	52.0	45.0	40.0

*Only the last attempt on each of the several exams is counted in the term mean.
**For 131S73 only, the minimum score needed to keep any grade was set at 35.0, and the minimum score needed to raise any grade was set at 75.0.

I. C. Evaluating the teaching method

The evaluation of the modular self-paced instructional system used in CEM 130/131 requires the measurement or estimation of several inputs and outcomes. The relevant outcomes of this system are the student's knowledge of chemistry and his attitudes about the course and subject. Measurement of these quantities is not merely clerical, but involves systematic investigation into the instruments of such measurement. As Havighurst [1] notes, "The key to improvement is very often the skillful devising of ways of observing and measuring achievements of the desired outcomes. For this reason the devising of tests and other instruments of evaluation is of great importance." The needed research in the areas of achievement and attitude is introduced in the next two sections. The third section treats the inputs of the system -- the costs in dollars, time, and effort.

I. C. 1. Knowledge gained by the students

Before knowledge of chemistry can be appraised, one must first decide exactly what to measure and how to measure it. Research into curriculum and course content is beyond the scope of these studies. The what of these courses is discussed in Chapter IV. In contrast, the details of how to measure achievement in this assigned body of chemical knowledge is eminently susceptible to systematic study.

I. C. 1. a. Research in measuring achievement

Research in measurement is an investigation of the errors inherent in different measuring instruments or techniques. Such errors in measuring student achievement are the topic of Chapter II. The goal of this research in measurement error is that through analysis and quantification, error can be predicted, controlled, and reduced.

I. C. 2. Student attitudes about the course and subject

The what and how of attitude measurement are less explicit than for achievement. Student self-reporting on attitude surveys is the institutionally accepted method of measurement. Specifically what attitudes are measured is discussed in Chapter IV.

I. C. 2. a. Research in measuring attitudes

Since individual grades are not assigned to students on the basis of their attitudes, there is less concern with individual random errors. What is still important, however, is the overall accuracy of group attitude measurement. The goal of this research in attitude measurement is to determine whether the survey techniques used in these courses are of the same relative accuracy as those used throughout the university. This experimental determination is described in Chapter III.

I. C. 3. Costs of the system

In an ideal learning system the students would learn more, like it better, and cost the department less. Unfortunately, it is not easy -- if indeed possible -- to achieve all three simultaneously. Some of one may have to be traded to gain more of another, and external conditions may impose maximum or minimum values on these variables, further limiting design choices.

As was stated earlier, one of the main assumptions upon which the new learning system is predicated is that certain costs must be decreased. The number of faculty and graduate assistants available for assignment to these courses may be less than in the past. Since personnel costs are the largest budget item, the more efficient use of fewer instructors may result in considerable saving. This saving will be offset at least somewhat by increased costs in computer charges, undergraduate labor, and equipment investments. The following compendium of the costs incurred in the operation of CEM 130/131 has been prepared. Some of these costs are adjusted averages and others are approximations, but most are compiled from the actual dollars spent by the Chemistry Department in the operation of these courses.

I. C. 3. a. Categories of cost

The most basic division of costs is into fixed and variable. The fixed costs of teaching a course remain substantially constant regardless of the enrollment. Some of these include maintenance, utilities, construction amortization, and administration. Since these costs are constant, and since the Department of Chemistry is not separately billed for these costs when a course is taught, such 'hidden' fixed costs will not be treated.

Costs which depend on the operation of a course, however they are related to enrollment, are treated as variable costs. Some are linear with respect to the number of students, such as the cost of handouts; some are step-functions, such as the cost of faculty and staff; and some are unrelated or related in a complex manner, such as development and improvement costs. The variable costs incurred by the Department of Chemistry have been divided into eight broad categories:

1. Faculty
2. Graduate assistants
3. Undergraduate assistants
4. Undergraduate hourly staff
5. Regular salaried staff
6. Computer programmers
7. Computer charges
8. Equipment, supplies and services

The definition of each of these categories is discussed below, and the extent to which quoted costs are approximate rather than actual is noted in the discussion of each category.

Faculty. In a typical lecture course, the faculty cost would follow a step function with each step equal to the capacity of the lecture hall. In CEM 130/131 there are no lectures, so this relationship is only approximate. Faculty are assigned to courses in the Chemistry Department in half-time units, of which there are seven per year.* Thus for any term, each unit costs one-seventh of a yearly salary. The mean salary for all faculty in the department was divided by seven for each academic year to arrive at the average cost per unit. The levels of half-time compensation for each unit of faculty for the relevant academic years are presented in Table 1.3 on the next page.

Graduate assistants. The cost of graduate teaching assistants is also a step function, but in contrast with faculty costs, the steps are much smaller and hence the relationship between cost and enrollment is smoother. Graduate teaching assistants are also assigned to courses in half-time units but the stipends paid to graduate assistants vary much less than do faculty salaries. There are three levels of appointment; the median stipend -- Level 2 -- has been chosen as representative of the cost of one graduate assistant unit. The change in average stipend over the years of these data is displayed in Table 1.3.

*Since summer session courses are often half-term, and duties during the summer are much lighter than during other terms, the faculty units for summer terms have been weighted only half as much as units from Fall, Winter, or Spring terms.

Undergraduate assistants. In Fall Term 1975, the Department of Chemistry began hiring undergraduate teaching assistants to work in CEM 130/131. Each undergraduate assistant is assigned in quarter-time units, thus making two undergraduate assistants equal to one graduate teaching assistant in the calculation of work load and the number of assistants needed. In Table 1.3 the average stipends paid to undergraduate assistants represent two quarter-time units in order that costs of faculty, graduate, and undergraduate assistants may be compared directly.

Table 1.3 Compensation in dollars per half-time unit paid to faculty, graduate assistants, and undergraduate assistants.

Year	Faculty	Graduate	Undergrad.
1972/73	2,667	1,170	-
1973/74	2,767	1,260	-
1974/75	3,058	1,350	-
1975/76	3,166	1,470	750
1976/77	3,296	1,530	780
1977/78	3,567	1,620	810

The cost per unit -- which represents what are assumed to be equivalent amounts of work -- decreases by half from one category to the next; a graduate assistant is slightly less than half as expensive as a faculty member, and an undergraduate teaching assistant is about

half as expensive as a graduate teaching assistant. These differences reflect the dissimilar duties and responsibilities of these positions.

Undergraduate hourly staff. Clerical and mechanical tasks such as exam grading, record keeping, and tape duplication are performed by undergraduates hired on an hourly basis. The actual number of hours worked and nominal wages paid are reported in the summary table in the next section. Almost all of the undergraduate student staff are paid through work-study grants subsidized outside the Department of Chemistry. The department pays only one-fifth of the nominal hourly wage of a work-study student. Also, a very small fraction of the total hours listed for undergraduate staff is attributable to record keeping work done for other courses. Thus the actual dollar cost to the department for undergraduate staff in CEM 130/131 is only twenty percent of the listed figures. However, subsidies of this nature are not universally available, so the more realistic nominal dollar cost for undergraduate staff is reported, rather than the actual dollar cost.

Regular salaried staff. Secretarial and supervisory personnel are not assigned specific course work-loads. The cost for staff members is an approximation based on the estimated percentage of their time spent on CEM 130/131, and as such may be in error ten or twenty percent. In November 1977 one full-time position devoted entirely to management of these courses was established.

Computer programmers. The total wages paid to hourly employed programmers does not reflect instructional costs related to enrollment in these chemistry courses. The computer programmers work almost exclusively on developing and improving the computer programs used in various functions within the system. Much of the developmental work is done during the summer when more time is available for implementation and testing of software changes. The reported figures are an overestimate because all developmental costs are charged to CEM 130/131 even though since the summer of 1973 other chemistry courses enjoy the use of the programs and capabilities thus made available. Yet the overestimate is not great since CEM 130/131 account for from 75 to 95 percent of the total computer utilization of the programs, and if not for the impetus toward their development provided by these courses, there may have been no programming work done.

Computer charges. There are two principal divisions of computer charges -- permanent file charges and computer time charges. The costs reported in the summary table are for those file and time charges attributable only to CEM 130/131 as estimated from records of the numbers of separate jobs run, files stored, and changes in dollar balances of different account numbers. In as much as the same accounts were used for daily course operation as well as for developmental programming, the two areas could not be separated; again this cost category is an overestimate of true operating cost and

is related to course enrollment in a complex and somewhat inverse manner. The percentage of computer charges directly credited to operation (as opposed to development) has varied unpredictably between twenty and eighty percent of the total reported computer charges. Without development there would be no operations; for this reason the costs of development and operating expense are not listed separately since the writing, testing, and changing of computer programs is an indirect but real operating cost. Developmental work is not proportional to course enrollment and does distort the total cost per term to a certain extent.

Equipment, supplies and services. This last category of cost is the sum of five related areas: specially rented equipment, routine copy charges, university printing services, amortized capital expenditures for equipment, and specially ordered supplies. Although several of these subtotals are estimated rather than exact quantities, a special effort was made not to under- or overestimate any specific cost and therefore the total cost for the five subtotals should be quite close to the true cost.

I. C. 3. b. Cost summary

The costs incurred by the Department of Chemistry in the operation of CEM 130/131 for each of the eight categories discussed in the previous section are presented in Table 1.5 on page 39. The total cost per term is listed in the second to last column and the cost per

student in the last column. Summer terms have been excluded from the cost-per-student column because of the pronounced distortion in total cost which is the result of developmental computer work done mostly during the summer.

The variations in cost per student are the result of fixed operating costs and step-function costs; the most cost-efficient terms are those of very large enrollment. Presented below in Table 1.4 are the total and average costs for calendar equivalent terms.

Table 1.4 Mean cost per student for five terms (1973-77)

Term	Enrollment	Total cost*	Cost per student	
			Total	Operating
Fall	8775	259,935.00	29.62	28.11
Winter	8813	246,397.00	27.96	26.85
Spring	3852	164,670.00	42.75	39.94
Summer	835	55,361.00**	66.30	60.31

*Included in these totals are all developmental costs incurred.
**Deleted from this total are all computer programming and user charges since these are disproportionately large in the summer. Operating costs do not include amortized capital expenditures nor developmental computer programming and use charges.

Spring and summer terms have higher relative costs because of fixed staff salaries and fixed costs included in supplies and services. These fixed costs, plus the step-function of faculty particularly distort the cost per student for summer terms. Even with the atypically high summer cost, these courses are still more cost efficient than nearly

Table 1.5 Summary of costs in eight categories for CEM 130/131

Term	N	Faculty Unit	Faculty $	Grad.Asst. Unit	Grad.Asst. $	Undergrad. Unit	Undergrad. $	Student staff	Staff	Prgmmrs.	Computer chgs.	Suppl. &serv.	Total cost	Cost/ student
F72		2	5,334	1	1,170				887.50			658	8,050	
W73	643	5	13,355	7	8,190			2,200	2,650.40	200	155.55	2225	28,976	45.06
S73	577	5	13,355	6	7,020			2,241.95	2,605.20	200.00	71.88	2123	27,617	47.86
SS73	215	1	2,667	2	2,340			373.20	2,510.40	1,224.00	104.54	1852	11,071	
F73	1312	3	8,301	15	18,900			2,193.85	4,460.00	395.78	353.23	3025	37,629	28.68
W74	2045	4	11,068	22	27,720			2,816.00	4,460.00	577.00	385.51	4139	51,166	25.02
S74	700	4	11,068	11	13,860			1,960.48	3,635.76	814.00	239.81	3111	34,689	49.56
SS74	153	1	2,767	2	2,520				4,323.65	1,170.00	337.81	2562	13,680	
F74	1445	2	6,116	14	18,900			2,363.18	4,510.40	442.53	585.61	3668	36,586	25.32
W75	1912	4	12,232	23	31,050			2,631.35	4,510.40	542.07	893.16	4016	55,875	29.22
S75	860	2	6,116	7.5	10,125			2,108.57	3,230.40	740.34	796.08	2781	25,897	30.11
SS75	165	1	3,058	2	2,700				2,862.40	1,740.23	506.85	1693	12,560	
F75	1772	2	6,332	15	22,050	5	1,950	3,708.21	3,424.90	648.96	908.29	3311	42,333	23.89
W76	2067	3	9,498	13	19,110	16	6,267	4,087.80	3,424.90	329.13	1104.89	4152	47,974	23.21
S76	918	2	6,332	11	16,170	10	3,942	3,707.31	3,424.90	485.81	1096.08	2498	37,656	41.02
SS76	186	1	3,166	0		1	1,294	332.48	2,987.20	1,949.76	1023.42	2018	12,771	
F76	2076	4	13,184	25.5	39,015	2	780	6,051.27	3,626.50	1,315.90	1302.30	3839	69,114	33.29
W77	2146	4	13,184	19	29,070	11	4,020	5,816.33	3,064.00	525.58	1690.58	5036	62,406	29.08
S77	797	2	6,592	11.5	17,595	5	1,908	3,946.83	3,064.00	348.04	1582.34	3775	38,811	48.70
SS77	116	1	3,296	2	3,060	0		1,732.22	3,191.20	1,879.71	1501.21	2055	16,715	
F77	2170	4	14,148	23.5	38,070	10	3,906	6,416.25	5,003.70	968.37	2048.59	3712	74,273	34.23

any other low-enrollment university course. For example, a class of twenty students with a full-time faculty member as the only item considered in the cost will have a cost per student more than twice as great.

I. C. 3. c. Comparison with the lecture method

In the previous two sections actual or absolute costs were discussed. In this section are presented the relative costs of instruction for equal numbers of students under the new modular self-paced method and the traditional lecture-recitation method. The operating costs over the past seven years during Fall and Winter terms are the basis of the figures presented in Table 1.6. The purely developmental costs of computer programming have not been included in the computer charges, and the total and per-student costs have been given both with and without amortized capital expenditures for equipment.

The cost per student for the lecture recitation method is almost independent of enrollment whereas that of the new method is strongly dependent on enrollment. The comparison of enrollments of one and two thousand students emphasizes the increased cost-effectiveness of the new method as enrollments increase. The new self-paced method has higher fixed costs for staff and equipment while the traditional lecture method has higher variable costs for instructional personnel.

Table 1.6 Comparison of operating costs in 1975 dollars for lecture (old) and self-paced (new) courses

Cost category	1000 students		2000 students	
	Old	New	Old	New
Faculty[1]	12,232	6,116	24,464	12,232
Graduate assistant[2,3]	12,150	12,150	25,650	24,300
Undergraduate assistant[3]	-	1,560	-	2,730
Student staff	-	1,488	-	2,976
Staff	1,644	4,417	1,644	4,417
Computer charges	106	210	212	420
Supplies & services	780	2,529	1,530	4,077
Total cost	26,912	28,470	53,500	51,152
Cost per student	26.91	28.47	26.75	25.58
Adjusted total cost[4]	26,912	27,070	53,500	49,752
Adjusted student cost[4]	26.91	27.07	26.75	24.88

[1]One unit of faculty per 250 students in lecture and one unit per 500 students in new method; based on seven year means for Fall and Winter terms.

[2]Seven year mean ratio of 108 students per T. A. in lecture course and 89 students per T. A. in self-paced method; figures rounded upward.

[3]Half-time appointments are divided eighty percent graduate and twenty percent undergraduate assistant.

[4]Adjusted total and per-student costs do not include amortized capital cost; only true operating costs are included in supplies and services for these figures. The amount deleted is $1,400 fixed cost.

As the enrollments increase, the significance of the fixed costs decrease and the new method becomes more dollar-efficient than the old method.

From a purely monetary standpoint the modular self-paced method of instruction now being used for CEM 130/131 is less expensive than a continuation of the lecture method would have been. In addition, there is one other major increase in efficiency which is not reflected in these figures. A graduate assistant in a lecture course is typically assigned five contact hours in recitation sections and help room. In CEM 130/131 the half-time assignment for each graduate teaching assistant is twelve hours in the help room with no other daily duties. Thus each teaching assistant provides more than twice the instructional output figured in contact hours for the same salary.

Chapter II Errors in measuring student achievement

II. A. Theory and literature

Quantitative analysis is not limited to the chemical laboratory, but occurs also in the classroom, for what is a chemistry test but the instrument with which the teacher estimates the different amounts of knowledge in each student in his 'sample'? The accuracy of a test score -- its goodness as a measure of the student's true level of knowledge -- is not susceptible of statistical error analysis. The systematic errors which affect the accuracy of the test are educed rationally or through some comparison with an external standard. A typical example of a systematic error is a test which is too easy or too difficult. This error is counteracted by shifting the grading scale upward or downward. The entire end-of-term process of constructing a scatter diagram of all students' scores, deciding on the average course grade, and choosing the cutoff points for each grade category is essentially a process of adjusting measurements for accuracy.

However, there may still be errors in measurement which do not affect the accuracy of the test score systematically but which do affect the precision of the test as a measuring instrument. A random or nonsystematic error lowers the precision or reproducibility of the individual scores often with no effect on the mean and distribution of all scores. This imprecision in the test score is recognized at least implicitly when grades are assigned to students 'on the borderline.'

Giving the student the benefit of the doubt or using additional criteria such as steady progress or evidence of effort are attempts to allow for the random errors in measurement.

In this chapter, errors in measurement which affect the precision of the test score will be explicitly examined both theoretically and experimentally.

II. A. 1. The probable error for a test score

The degree of reproducibility or 'probable error' of a measurment is specified by a plus-or-minus interval around the observation. The probable error is a function of the scatter in the experimental data expressed as the standard deviation of all observations from their mean [31]. Thus the precision of a measurement is estimated from the variation among several measurements.

The probable error or standard error of measurement (SEM) of an analytical balance could be determined by selecting a typical mass and making several weighings. The assumptions are made that the mass is unchanging and that once calibrated for this mass, the balance may be used to estimate masses at other points in the scale with equal precision. By analogy, the amount of chemistry a student knows might be considered as his 'mass of knowledge' and the examination as the balance we use to estimate this mass. However, there are three reasons why we cannot determine the standard error for a test by simply choosing a student and administering the test a dozen times.

First, the same student if given the same test a second time would remember many of the answers he gave the first time if the interval between testings were short, or change in the amount of chemistry he knew if the interval were long. Second, the scale of test scores, unlike the mass scale of a balance, is not a linear interval scale. A zero on the test does not mean that a student knows no chemistry, nor does a score twice as large as another imply twice as much knowledge. Third, the teacher is rarely interested in the precision of one specific test since seldom is the same test used more than once. Rather he is interested in the precision of a class of tests which might be generated according to a specific procedure. For these reasons we must attempt to redefine the standard error of measurement in terms both manageable and meaningful. A definition that is manageable will define the precision of a test in variables which can be estimated with reasonable confidence given the characteristics of the students and the tests. A definition that is meaningful will express precision in terms which are related to the quality of the test.

II. A. 2. The importance of test reliability

The standard error of measurement, SEM, for a test score may be derived from the binomial formula for sampling error.* The equation is

$$SEM = \sqrt{k\,p\,q} \qquad [2.1]$$

where k is the length of the test, p is the true fraction of the material known by the student, and q is the fraction unknown. An unbiased estimator of the standard error is

$$SEM \simeq \sqrt{\frac{1}{k-1}(x)(k-x)} \qquad [2.2]$$

where x is the observed test score [33]. Two conclusions can be drawn: a longer test provides a more precise estimate of the student's knowledge, and the measurement error depends on the score the student gets. Yet although manageable, this definition is not meaningful in the desired sense since neither test length nor test score are related to test quality.

*Lord [32] compares the student to the statician's urn of black and white balls: "Paralleling the urn containing a large number of balls, we may imagine a pool containing a large number of test items. If all the items in the pool could be administered to the examinee without practice effect, fatigue effect, and so forth, the ones he would get right may be thought of as corresponding to the white balls, and the ones he would get wrong as corresponding to the black balls. Each parallel form of the test is thought of as a random sample from the pool. The number of 'white' (correctly answered) items in each sample is the examinee's score on that form of the test. The standard deviation of this number, found by the usual binomial formula, is the examinee's standard error of measurement..."

Lord [34] has shown that the mean standard error as defined in Equation 2.2 is related to test reliability by

$$\overline{SEM} = s_x\sqrt{1 - r} \qquad [2.3]$$

where s_x is the observed standard deviation and r the reliability calculated according to the Kuder-Richardson Formula 21. This relation, Lord states, "...is a mathematical identity -- not an approximation, not a conclusion based on plausible assumptions." The average standard error (a geometric mean) is defined by this equation as a function of test reliability. This definition is somewhat less manageable than the ones described by Equations 2.1 and 2.2 because estimates of test reliability are less determinate than are test length and test score; but it is more meaningful because test reliability can be clearly related to test quality. The focus of this discussion will shift to the definition and estimation of test reliability.

II. A. 3. Definition of test reliability

The test score in classical test theory is defined as the sum of true score and error [35]. The true score is the score the student would receive if there were no errors of any kind in measurement. Error is a random contribution which may make the observed score higher or lower than the student's true score. A collection of test scores will have a separate mean and variance for true scores and for error scores. The mean of the true scores is the 'true test mean.'

Since error scores are assumed to be purely random, the mean of all error scores is zero.* The variance of the true scores is positive because there are real differences in knowledge of chemistry between students. The variance of error scores is positive because squared deviations from a mean of zero do not cancel.

There may be many independent causes of random error, each contributing some fraction to the error variance. The observed test variance will be the sum of the true-score variance and all the error variances.** A typical breakdown of test variance into its components is listed in Table 2.1.

The reliability of a test is formally defined as the fraction of total variance produced by differences in true scores among students. Since the true score is unknown, reliability cannot be definitely determined; it can only be estimated. The choice of a statistic with which to make this estimation will depend on the assumptions one

*This conception of true and error scores is a simplification. Error may be systematic or random. Systematic error may be either constant for all scores, increasing or decreasing each by the same amount, or it may vary in a complex manner depending on test score or some other variable. Systematic error affects the absolute accuracy of the scale without affecting relative accuracy or ranking. Random error directly affects the precision of the score -- the confidence with which a higher score is held to indicate greater knowledge than a lower score. Random error cannot be 'corrected for' the way one corrects for systematic error; it can only be estimated after-the-fact. Through post hoc error analysis the sources and sizes of random error can be determined and perhaps attenuated in future measurements. In discussions of precision and error in measurement, systematic error often (as here) will be included with true score so as to simplify the analysis of random error.

**The variance of a sum of uncorrelated scores is the simple sum of the individual variances of the separate score distributions.

chooses to make about the students and tests involved. Several methods of estimating test reliability will now be discussed with attention given to how different assumptions about true and error variances affect the interpretation and use of the coefficient thus estimated.

Table 2.1 Components of observed test variance

Symbol	Source
s^2_T	differences in knowledge (T = true score)
s^2_{fl}	learning and forgetting (fl = fluctuation in true score)
s^2_g	guessing variations among students (g = guessing)
s^2_{adm}	effects of day, room, etc. (adm = administration)
s^2_{eq}	unrepresentative coverage of topics (eq = equivalence)
s^2_{sc}	subjective scoring errors (sc = scoring)
s^2_m	memory of previous test questions (m = memory)
s^2_R	all remaining sources of error (R = residual)

II. A. 4. Reliability from test-retest

The repitition of a measurement and the correlation between these repeated measures as a way to estimate the reliability involves only two assumptions: (1) the quantity measured is unchanged by the act of measurement, and (2) the quantity measured does not fluctuate during or between measurements. While these assumptions may often

be met in the chemical laboratory, it is unlikely that they are both true simultaneously in the classroom. The smaller the interval between administrations of the same test, the less likely true scores fluctuate but the more likely memory of the first set of responses influences the subsequent set. The score the student gets may be the same not because his knowledge of chemistry is unchanged but because he remembered how he answered those questions the first time. This inclusion of memory variance with true-score variance will spuriously inflate the estimate of the reliability coefficient [36, 37]. If sufficient time is allowed between tests for the students to forget how they answered the questions then they probably have also forgotten some of what they knew. The reliability coefficient may also be inflated when calculated by the test-retest method if the specific test used is not representative of other possible tests. This second type of error, commonly called sampling error, is not important when the property being measured is very narrowly defined. For example, whether a particular yardstick is 'representative' of possible yardsticks is easily judged. Achievement in chemistry is not as easy to measure accurately as is length or the yield of a chemical reaction.

II. A. 5. Reliability from parallel forms

The estimation of reliability from the correlation between two parallel forms of a test is a technique which again involves only two assumptions: (1) two 'different' tests are 'equivalent', and

(2) there is no fluctuation in the student's knowledge during or between tests. The second of these assumptions is the same as the second upon which test-retest reliability measurements are based. The basis of the first assumption is the assertion that two tests have the same content. The definitions of equivalence range from item-by-item correspondence between forms [38-40] to sets of items randomly selected from a pool of parallel items [41-43]. The product-moment correlation between these two tests when given to the same group of students is the geometric mean of their reliability coefficients [44]. Much if not all of the variance due to sampling error is transferred from true score to error, and memory as a source has been eliminated. The parallel forms method of reliability estimation is the accepted operational definition of test reliability.

Estimation of reliability by either test-retest or parallel forms correlation still involves the assumption about true-score fluctuation over a time interval. Cureton [40] says, "Ideally, test reliability should be determined from experimentally independent test sessions so close together that the true abilities of the examinees do not change during the interval. In practice no interval is short enough." It would take five administrations at equal intervals to provide a reasonable extrapolation back to an interval of zero [44]. Because of the extra labor involved in the construction and administration of more than one test, a statistic which estimates the reliability from only a single test administration is often used.

II. A. 6. Reliability from homogeneity of content

Two methods based on the internal consistency of a single test may be used to estimate the reliability coefficient: the split-halves correlation pioneered by Spearman in 1910 [35] and the types of coefficients introduced by Kuder and Richardson in 1937 [38]. When a test is split into two halves the assumption is made that since the whole test was taken at one sitting there is no time interval between halves.* If each of these halves is fully representative of the whole test then the correlation coefficient between them (adjusted upward to the length of the whole test using the Spearman-Brown prophecy formula) will be as appropriate an estimate of test reliability as would be a correlation between parallel forms.

Coefficients of internal consistency make the assumption that within-test item correlations are as appropriate as between-test correlations. This minimum assumption is the basis for the Kuder-Richardson Formula 8 (KR_8). In the derivation of the KR_8 it is assumed that item-test correlations, r_{it}, are the basis for reasonable substitutes for item reliabilities, r_{ii}, via the relation

$$r_{it} = \sqrt{r_{tt}\, r_{ii}} \qquad [2.4]$$

*According to Cureton [40] a splitting of the test into two halves when the number of items is a multiple of four may be done thus:

1 4 5 8 9 12 13 16
2 3 6 7 10 11 14 15

The midpoints in the administrations of the two halves coincide for this pattern. In the usual odd-even split, the 'even' half is administered a full item later than the 'odd' half. This time displacement is probably inconsequential unless the individual items are lengthy.

If in addition it is assumed that items are equally intercorrelated and have equal variances, the KR_{20} may be used to estimate consistency. By making a final rigid assumption that all item difficulties are equal,* internal consistency may be estimated from only the test mean and test variance by the KR_{21}.

II. A. 7. Factors affecting test reliability

The assumptions one is willing to make influence the choice of a statistic to be used as the estimator of reliability. It is therefore desireable to know how violation of these assumptions affect the estimate. Many authors have examined the interrelations of test length (k), test variance (s^2), item variance (pq), mean item difficulty ($\bar{p}$), number of answer choices (J), item intercorrelations (r_{ij}), distribution of item difficulties, factor structure of test content, internal consistency, and reliability.

II. A. 7. a. Test length

A longer test is a more reliable test. This axiomatic relationship was first quantified by Spearman and Brown separately in 1910 and is expressed by the Spearman-Brown prophecy formula

*This assumption is rarely met but is seldom critical. If item difficulty is 0.50, item variance is 0.25; as the difficulty diverges to 0.36 or 0.64 the variance decreases only slightly to 0.23. Thus only when item difficulties are widely distributed about their mean does the average item variance differ significantly from the square of the average item difficulty.

$$R_n = \frac{n\,r}{1 + (n - 1)\,r} \qquad [2.5]$$

where r is the reliability of a test and n is the factor by which the length of the test is decreased or increased. This formula is derived from classical test and information theory which considers the 'true' component of test variance as increasing with the square of n while the random component increases only linearly [35, 45, 46]. The Spearman-Brown prophecy formula is based on the assumptions that variances and reliabilities per-unit-length added or deleted are equal. Cureton [40] demonstrated that the violation of either assumption separately will have little effect on the prediction. (This demonstration provides post hoc justification for what would otherwise throughout the years have been rather indiscriminate application by many who discuss predictions but not assumptions.)

II. A. 7. b. Number of response options

The effect of the test length on reliability has been discussed in the previous section. Yet the definition of the length of a test as the number of items is but one of two limiting-case views of how to calculate the length of a test. The first case considers the length of the test to be the total number of answer choices for all test questions. In this case, the item is treated as an inconsequentially short stem plus a collection of response options from which the correct answer must be selected upon careful consideration of all answers in the set.

Thus the student spends his time selecting his response, and a better representation of the test length is the total number of response options rather than simply the number of questions.

The second limiting case consists of a problem to be solved without reference to a set of responses. Upon solution of the problem, the answer need only be located quickly in the set of answer choices. The student's time in this case is almost wholly spent solving a problem instead of making a choice. The reliability of these stem-content items should follow the Spearman-Brown prophecy formula with the test length conventionally defined as the number of questions. The reliability coefficients for response-content (first case) items should follow the prophecy formula when the test length is taken as the total number of answer choices. Several authors have studied the prophecy of reliability for what they assume are response-content test items.

Remmers and co-workers [47-52] reported several comparisons of split-halves correlations with reliabilities predicted by the prophecy formula. The formula correctly predicted reliability coefficients for items with two through five choices taken from the Purdue Placement Test in English and prepared by random deletion of incorrect options from the original five-choice questions. In another study of arithmetic items for junior high school students, he found again that the predicted reliabilities agreed with experimental values. A third similar study confirmed the prophecy involving algebra tests of varying numbers of

answer choices. Remmers did not examine the validity of the prophecy formula for achievement tests containing questions with more than five response options. However, he did observe that the formula failed to predict the reliabilities of seven-choice attitude survey items.

The Spearman-Brown prophecy formula predicts that a test of infinite length will have a reliability coefficient of 1.00 irrespective of the initial reliability upon which the prophecy is based. When the test length is considered as the total number of choices, the short-answer or free response item is a special case of the item with an infinite number of choices.* This leads to the conclusion that all short-answer tests are of perfect reliability simply because of their structure and independent of their content or quality. This seems unlikely. What has been supported by Remmers is the general conclusion that increasing the number of answer choices does increase reliability. The exact point at which to expect breakdown of the Spearman-Brown prophecy formula has not been determined. It is interesting to note that even for supposedly stem-content arithmetic and algebra questions the formula correctly predicts increased reliability for up to five options; this is more likely due to a decrease in guessing than to an increase in apparent test length.

If the total number of options -- test length -- is held constant, the interaction between the specific number of questions and the

*This is done explicitly by Mattson [53] and implicitly by other authors.

probability of a successful guess will lead to an optimal number of response options per item. Tversky [54] calculated the theoretical maximum for a discrimination function. The optimal number of options was the transcendental constant e (2.718...), which rounds up to the integer three.

Grier [55] used an equation developed by Ebel* to show that the KR_{21} is also maximized at three options per item. He stated three reservations. This analysis is only valid for response-content items. It applies only to tests where the total number of options is greater than fifty-four. And finally, only when the use of the KR_{21} is a reasonable basis for estimating test reliability are these conclusions appropriate.

More recently, Lord [57] analyzed the results of the SAT Verbal exam. With the assumption that guessing is purely random, he derived an equation to predict the number of choices which maximizes the reliability coefficient independent of test length but dependent on item intercorrelations and difficulties. This equation yielded predictions similar to those of Tversky and Grier. However, Lord found that the relative efficiency of items with different numbers of choices is a function of the difficulty level. At low ability (high difficulty), guessing is of paramount importance and he found that five-choice items were

*Ebel's [56] equation is based on reasonable assumptions about the test mean and standard deviation; it provides a lower-bound estimate to the reliability. The estimate becomes increasingly less accurate as test length decreases or as the standard deviation increases.

more effective than two-, three-, or four-choice items. Conversely, at high ability (low difficulty), reliability and discrimination depend on the number of items -- to attain the maximum number of items for a given total number of options requires the fewest (two) per item.

By adjusting item difficulties according to suggestions Lord made previously [58], he eliminated much of the effect of differential ability (difficulty level) and found the three-choice format to be most efficient. Too, the predicted chance score based on the number of response options often overestimates the actual chance score. Lord observed that low-level examinees perform at below chance. Lord also added that seldom is a guess truly random, and when the response is based on partial information, the discriminating power of the item is likely to change in an unpredictable manner when the number of answer choices is changed. This tends to discourage any attempt to restructure items by lowering the number of answer choices in order to increase the number of items.

Costin [59] did an empirical study of the effect of reducing the number of response options from four to three as a test of Tversky's proof that three is the optimal number. He selected fifty to sixty items from each of four topics in an extensive pool of items used in an introductory psychology course. Half of the four-choice questions were reduced to three-choice by random deletion of a distractor. After two hundred students took the tests, the items were sorted into the two types. The discrimination index, KR_{20}, and item difficulties

were calculated based on test lengths of twenty-five or thirty items. While the overall values of the KR_{20} coefficients were not impressive (0.50 to 0.62), for all topics the three-choice items were more consistent, more discriminating, and only slightly less difficult. Though many of the differences were not statistically significant, the trend in the coefficients supported Tversky.

Ramos [60] distinguished between 'natural' four-choice items and 'artificial' four-choice items. Artificial four-choice items were originally five-choice items from which the least attractive distractor was deleted. College-level French and Spanish reading tests were administered to over one thousand students at different schools. After correcting scores for guessing* and using a KR_{20} modified for omitted items, Ramos found that artificial four-choice items had coefficients comparable to natural four-choice items. He also found that natural four-choice items had slightly lower internal consistencies than the unmodified five-choice test items. Ramos pointed out that generalizing to other dissimilar tests is not entirely warranted.

These studies of predicted reliability and the optimal number of answer choices are based primarily on response-content items in areas other than chemistry. Most of the questions asked on chemistry

*When test scores are corrected for guessing, Zimmerman [61] and Coulter [62] agree that the average effect of guessing on the group mean may be eliminated, but the irregular and unknown effects of luck for different students will still contribute to error variance. Coulter points out that if all items are attempted, corrected scores correlate perfectly with uncorrected scores.

tests in CEM 130/131 are stem-content items such as numerical problems. The length of time it takes a student to work a problem is theoretically independent of the number of answer choices. Thus problems are not susceptible to the same mathematical treatment by which three is determined to be the optimal number of choices. Yet some points made by Remmers and Lord are pertinent. The reliability of tests will increase when the number of response options is increased. This increase will be most effective when the tests are somewhat difficult, and most of the gain in reliability will occur at the low end of the scale.

None of these studies went beyond five answer choices. It is a major aim of this research to extend the investigation of reliability into the uncharted sea between five and ten response options.

II. A. 7. c. Interrelation of topics

The manner in which items testing the various topics on an examination interrelate is described by the factor structure from a factor analysis. When the coefficient of internal consistency is used as an estimate of test reliability, the assumption is made that item intercorrelations are homogeneous.* A reliability estimate 'half way' between internal consistency and parallel forms correlation is the

*This assumption of homogeneity does not mean that all items must intercorrelate highly or even equally. It does mean that there are no clusters of items which intercorrelate more highly among themselves than they do with other items on the test.

split-halves correlation coefficient. Cronbach [43] compared the KR_{20} to the split-halves coefficient. He showed that the mean of all possible split-halves correlations is equal to the KR_{20}. Given the logical presumption that a planned division would produce a higher coefficient than the mean of all possible splits -- even poor ones -- Cronbach analyzed a 60-question test of mechanical reasoning taken by ninth grade boys. Three schemes of splitting the test into two halves were employed: random selection, selection according to item difficulty, and selection according to both content and difficulty. The range of calculated coefficients was greatest for the random splits and smallest for splits equated for both content and difficulty. The mean of all possible splits calculated by the KR_{20} was roughly equal to the mean of the random splits and slightly lower than the planned splits, but the differences were not statistically significant. Referring to other analyses, Cronbach added "...we have studies of seven tests which seem to show that the variation from split to split is too small to be of practical importance." This failure of seemingly heterogeneous tests to have lower coefficients of consistency than reliability led him to study the influence of factor structure on internal consistency.

Three basic kinds of factors are found in test items -- a general factor (G), group or cluster factors, and single-item factors. The sizes and distributions of these factors within the test have differing effects on consistency. If there is only one significant general factor

and no group or cluster factors on a test, then even when this general factor is weak (i.e. item intercorrelations of 0.09) its importance rapidly increases with test length. For a test of twenty-five items, G accounts for about 45 percent of test variance and for one hundred items, G accounts for about 75 percent. In this case where there are no distinct clusters or subsets of items, then as long as the test is not simultaneously short and heterogeneous, the internal consistency will be nearly identical to the split-halves correlation and by analogy the parallel forms correlation.

The KR_{20} indicates how much variance depends on the general factor. As such it is a measure of content homogeneity as well as test reliability. If a test is known to be heterogeneous with discrete homogeneous subtests, the concept of overall consistency does not apply and a coefficient of internal consistency will seriously underestimate the reliability. In summary, Cronbach stated that the KR_{20} is a lower bound to the reliability coefficient and an upper bound to internal consistency. The extent to which the KR_{20} underestimates reliability or overestimates consistency depends on the factor structure and the size of the G factor.

Lord [63] shows that estimates of internal consistency provided by the KR_{20} and KR_{21} are estimates of the random sampling error in the selection of test items. Consistency coefficients treat differences in topic areas and item content and difficulty within tests as the source of differences between tests. But if the pool of items from which those

on the test are selected is carefully stratified, it is not difficult to generate heterogeneous tests which are still parallel. In this case the parallel forms correlation coefficient will provide the only realistic estimate of true test reliability.

II. A. 7. d. Item intercorrelations and difficulties

Gulliksen [64] investigated the relation of item intercorrelations and difficulties to test reliability. By defining the reliability as the ratio of true variance to observed variance, he showed what conditions maximize test variance. For a well-constructed one-factor test consisting of items with similar reliabilities and intercorrelations which are also uncorrelated with item variance (item difficulty), he summarized the results in three theorems.

Theorem A. Item intercorrelations can be at a maximum only when item difficulties are equal, and the farther from 0.50 the average difficulty is, the faster correlation decreases as the item difficulties become more widely scattered about their mean.

Theorem B. Test variance will increase as item difficulties cluster more closely, as intercorrelations increase, and as the average item difficulty approaches 0.50.

Theorem C. As the test becomes more homogeneous in content and the items more similar in difficulty (ideally all equal to 0.50), the test reliability will increase. This is a corollary to the relation between test length and reliability.

Gulliksen points out that his theoretical treatment does not consider items on which guessing can occur. He also states that test variance is not the only criterion by which test reliability can be judged.

Brogden [65] studied the effects of variations in item difficulty, spread of item difficulties, and item intercorrelations on coefficients of internal consistency. In a simulation, he found that as item intercorrelations decrease, indicating an increasingly complex factor structure, the variance of item difficulties has a greater effect on consistency. This effect increases as the average difficulty departs from 0.50. Thus if items have low intercorrelations and widely different difficulties, even very long tests will have low internal consistency as estimated by the KR_8. With heterogeneous content there is more to ge gained by clustering item difficulties near the middle of the difficulty scale than with homogeneous content. The coefficients of consistency thus indicate not only consistency of content but also consistenty of item difficulties, and parallel forms of tests must be parallel in both content and difficulty.

Cronbach [66] discusses the relationship between spread of item difficulty and discriminating power at different levels of achievement. He demonstraes that the validity estimated by the biserial item-test correlation, r_{it}, is dramatically affected by the interactions between item intercorrelation and difficulty level. The optimal spread of item difficulties which provides the widest range of discrimination and validity depends on how highly items intercorrelate. When items are

very highly intercorrelated, the test should have a wide range of item difficulties so that discrimination is validly made at many levels of achievement. When item intercorrelations are low, which for a one factor test means low item reliability, the item difficulties should all be equally difficult at or just above the midpoint between chance and perfect.

II. A. 7. e. Guessing, difficulty, and consistency

Lord [58] extended the calculations of Brogden and Gulliksen to the case where guessing can occur. Using a KR_8 modified for guessing, he concluded that the internal consistency coefficient is maximized by decreasing the spread of item difficulties and by making the mean item difficulty slightly greater than halfway between chance success and 1.0. Lord noted that the consistencies are lowest fortwo-choice items and increase steadily as the number of choices increases to five per item. He also observed that the amount by which items should be made easier to maximize consistency increases as the intercorrelations among the items increase. He states that making items easier increases reliability at the low end of the scale while slightly decreasing the discrimination at the upper end.

Carroll [67] also extended Gulliksen's work to the case where guessing can occur. He derived similar relationships and reached similar conclusions. The reliability as estimated by the correlation between tests or between items decreases as the mean difficulty

increases and also as the differences in difficulty between tests or items become more pronounced. Carroll states a formula for estimating the true correlation after deletion of guessing variance

$$r' = \frac{r\, s_a\, s_b}{\sqrt{s_a^2\, s_b^2 - \frac{1}{J} E_a\, s_b^2 - \frac{1}{J} E_b\, s_a^2 + \frac{1}{J^2} E_a\, E_b}} \qquad [2.6]$$

where E is the mean failure score and J is the number of answer choices. When the two forms have equal means and variances, the formula simplifies to

$$r' = \frac{r\, s^2}{s^2 - \frac{1}{J} E} \qquad [2.7]$$

Nunnally [68] also derived a formula for correcting reliability coefficients for guessing. He began with the classical definition of reliability as the ratio of true to observed variance, and assumed that guessing not only increases error variance but also decreases true-score variance. His formula is

$$r' = \frac{2J - 1}{J^2} + \frac{(J - 1)^2}{J^2} \left[\frac{r\, s^2}{s^2 - \frac{1}{J} E} \right] \qquad [2.8]$$

The term in brackets is Carroll's simplified formula and the other terms are corrections for the decrease in true-score variance. Part of the guessing variance is related in a complex manner to true-score variance because guessing is greatest at low ability levels and decreases as true score increases. Carroll's correction formula is

limited to the product-moment correlation coefficient; Nunnally's correction formula may be applied to any reliability coefficient.

II. A. 7. e. Distributions of scores

While it is often assumed that the underlying ability measured by a test is normally distributed, the actual distribution of test scores may be quite unlike the normal curve. Scott [69] defined a formula for discrimination related to the distribution of scores and examined the effects of variations in test length, item intercorrelations, and difficulty on his index of discrimination. He found that maximum discrimination occurs when item difficulty is 0.50 and the observed distribution is rectangular. (The item intercorrelation describes the shape of the distribution -- when the intercorrelation is 0.00 the distribution is normal, when the intercorrelation is approximately 0.33 the distribution is rectangular, and as the intercorrelation nears 1.00 the distribution approaches a U- or J-shape.) As the test difficulty departs from 0.50 the maximum discrimination occurs when items are less intercorrelated than 0.33. Increasing test length so as to increase discrimination and precision is most effective when the item intercorrelation is at its 'best' value for a particular difficulty level. As the intercorrelation departs in either direction from its calculated optimum, the incremental improvement in discrimination gained by lengthening the test decreases. From empirical evidence he collected, Scott concluded that allowing the item difficulties to vary

widely about their mean value is not a dependable way to increase discrimination even when item intercorrelations are high.

Lord [42] considered some of these same interactions for the case of a one-factor test measuring a normally distributed ability. He drew several conclusions. "The test score distribution will not in general have the same shape as the distribution of ability; in particular if ability is normally distributed, the raw scores in general will not be normally distributed." "Typically, if a test is at the appropriate difficulty level for the group tested, the more discriminating the test, the more platykurtic [flatter] the score distribution." "The skewness of the test score distribution typically tends to become positive as the test difficulty is increased, negative as the difficulty is decreased." "U-shaped and rectangular distributions of raw scores can be obtained if sufficiently discriminating test items can be found..." or if the test is administered to a group with wide ranging abilities. "A test composed of items of equal discriminating power but of varying difficulty will not be as discriminating in the neighborhood of any single ability level as would a test composed of similar items all of appropriate difficulty for that level." Lord also notes that the distribution of error varies with ability level and "...although uncorrelated with ability and with true score in the product-moment sense, the errors of measurement are not independent of ability or of true scores, since the standard deviation and the skewness of the errors vary with the ability level."

II. A. 8. Factors affecting the standard error

This discussion thus far has focused on test reliability because test reliability can be directly related to important sources of error in measurement. In addition, test reliability has been examined as a function of such variables as test length, number of choices per item, and factor structure. The reliability, test variance, and standard error of measurement (SEM) are related as shown by Equation 2.3, restated here

$$\overline{SEM} = s_x \sqrt{1 - r} \qquad [2.3]$$

It is through changes in the reliability coefficient and the standard deviation of the test scores (s_x) that changes in the standard error of measurement are observed. Some of these changes will now be discussed.

Zimmerman [61] discussed the relation of reliability and standard error of measurement to guessing. The standard error of measurement changes with the opportunity for guessing. At higher test scores, it is less likely guessing is significant and this lowers the standard error. At low scores, guessing is more likely and hence the standard error is greater. Zimmerman stated a formula for calculating the minimum possible standard error based on the number of answer choices (J), the test length (k), and the test score (x).

$$SEM_{min} = \sqrt{\frac{1}{J}(k - x)} \qquad [2.9]$$

This residual error (SEM_{min}) would be present even if all other sources of error were absent. The minimum error decreases as the number of answer choices increases and as the true proportion of the material known by the student increases.

Horn [70] derived four different equations for the standard error of measurement which depend on different theories of error variance and reliability estimation but are calculated from the same summary statistics (observed test variance and parallel forms correlation coefficients). Equation 2.3 (restated below) applies when the true coefficient is estimated by placing a confidence interval around the observed coefficient. When two fallible measurements are compared, Equation 2.10 could be applied.

$$SEM = s_t \sqrt{1 - r} \qquad [2.3]$$

$$SEM = s_t \sqrt{2(1 - r)} \qquad [2.10]$$

Horn described two other formulas based on regression models. He recommends that the choice from among these four equations be appropriate to the variance model chosen rather than to the method of reliability calculation. Since the variance model most commonly chosen treats the calculated reliability coefficient as the best estimate of true reliability, Equation 2.3 is typically used.

Mattson [53] pursues a derivation of the standard error of measurement which shows a relation to difficulty and reliability

somewhat different from that discussed by Zimmerman. On a difficult test (proportion correct less than 0.50) guessing will increase the standard error of measurement, whereas on an easy test guessing will decrease the standard error. For any difficulty level, an increase in the chance of a correct guess lowers true-score variance and hence reliability.

The formulas presented above for the computation of the standard error of measurement pertain to a single test. A series of tests given to groups of students may be considered as a set of t subtests of k items each from a pool of K items, and the students as subgroups of n students from a population of N students. Shoemaker [71] described a formula for the standard error of the total mean score of all students on all items based on a multiple matrix model of sampling subgroups of students and subtests of items. The formula for an infinite population of students is

$$SEM = \sqrt{\left(\frac{1}{tkn}\right)\left(\frac{1}{k-1}\right)\left\{\frac{K^2 s_d^2\left[(K-k)(n-1)-kn(t-1)\right]}{+K^2 s^2(k-1)+\hat{\mu}(K-\hat{\mu})(K-k)}\right\}} \quad [2.11]$$

where s_d^2 is the variance of item difficulties, s^2 is the test variance, and $\hat{\mu}$ is the expected population mean for the entire pool of test items. Shoemaker indicated three relationships which follow directly from the formula. First, increasing the spread of item difficulties increases the standard error of measurement unless $tk = K$. Second, increasing

internal consistency (as measured by the KR_{20}) increases the standard error. Third, increased skewness in the distribution of scores also increases the standard error of measurement. Shoemaker empirically examined the size of the expected effects through a series of computer-simulated tests and found that consistency and spread of difficulties are important factors in the model. He also noted that increases in the number of students taking a subtest, the number of subtests, and the length of the subtests also reduce the standard error to varying degrees. The effect of the number of subtests is greatest for skewed distributions, and the effect of the test length greatest for normal distributions. Unless $tk = K$, the best general procedure is to increase the number of subtests rather than increasing test length.

This procedure may be applied to estimation of the mean score for a pool of test questions which cannot all be given to the population of students. The assumptions on which the model is based include one which requires subgroups of students tested to be randomly selected without replacement. The effect of overlap between subgroups of students was not discussed by Shoemaker but the effect would seem to make the estimated standard error of measurement an underestimate.

II. A. 9. An experimental comparison of two item formats

Plumlee [72] tested thetheoretical predictions of the effect of chance success on item-test correlation (an estimate of one-factor validity) and on test reliability. She compared five-option multiple-

choice items to short-answer items on a test of algebra, geometry, and trigonometry. According to theory, the reliability of multiple-choice items will always be less than the reliability of short-answer items because guessing imparts additional error variance to the observed variance. Since the true-score variance is at best unchanged, the ratio of true to error variance will be lower for the multiple-choice format. Also, the mean of a multiple-choice test should be higher than the mean of an equivalent short-answer test.

The expected mean of the multiple-choice test can be predicted from the short-answer test mean by

$$\bar{p}' = \frac{1}{J} + \frac{J - 1}{J}(\bar{p}) \qquad [2.12]$$

where J is the number of choices (in this case J = 5). Four sets of thirty-six items were prepared so as to be equivalent in content, difficulty, and discrimination. Each of the four sections was prepared in both short-answer and multiple-choice format. Two sets of each format were compiled to make two different test pairs, and each test pair was given in both directions -- multiple-choice followed by short-answer to one group and the reverse to another group. The four sections of thirty-six questions were preceeded by a common set of sixteen items in order that group equivalence might be estimated. The four batteries were administered to approximately 560 examinees (N = 139 for each battery) as separately timed subtests at one sitting. Her results are presented in Table 2.2 on the following page.

Table 2.2 Summary data from Plumlee [72]

	Group W	Group X	Group Y	Group Z
μ	37.7	39.8*	36.2	42.2*
σ	11.3	10.4	10.9	10.6
μ	37.8*	31.6	39.7*	31.1
σ	10.2	10.6	9.6	10.3
r(SA)**	0.84	0.80	0.85	0.81
r(MC)	0.81	0.76	0.72	0.81
$\hat{r}$	0.75	0.69	0.75	0.69

*Test means indicated with an asterisk are for multiple-choice forms; unstarred test means are for short-answer forms.
**SA indicates short-answer and MC indicates multiple-choice; r-hat is the multiple-choice coefficient predicted from the short-answer result.

Plumlee calculated the means and standard deviations based on the sum of the two sections of each test which had the same format. The reliability coefficients are parallel forms correlations between the two 36-item subtests in each battery that were of the same format and taken by the same students. Plumlee stated that the means on the common 16-item subtest used to equate the groups were statistically equal though she did not report them. A chi-square test for equality of reliability coefficients demonstrated that the differences among the observed coefficients were not significant.

Plumlee analyzed the regression prediction for test means and found that seven of the eight observed differences between sets of items with the same content but different formats were within the range

of the predicted values. However, these comparisons were for tests with parallel content administered to different groups of students. Inasmuch as the groups were equated by a test with only sixteen items, and for an alpha level of only 0.05, any conclusions are tenuous.

Plumlee herself gave further caveat by stating that the number of students who completed items in different sections of the battery ranged from 32 to 109 and thus the proportions correct ($\bar{p}$) used in the regression analysis are not based on N = 139. The high number of omitted items and the lack of control for this factor affects the strength of her conclusions.

Also listed in the table are the expected correlations of the multiple-choice tests as calculated according to Carroll's formula (Equation 2.6). The differences between the expected and obtained coefficients were statistically significant only for test Z.

Plumlee concluded that there seemed to be a small but consistent (although not statistically significant) tendency for the actual reliability and mean of a multiple-choice test to depart from the predicted value in the direction of the short-answer statistic. "...the evidence does seem to indicate that item-test correlation and test reliability may not be as adversely affected by the multiple-choice form as has been frequently assumed." The direction of this departure agrees with Lord's [57] observation -- examinees who do not know the correct answer respond at a below chance level and this tends to reduce the contribution of guessing to error variance.

Table 2.3 was prepared by this author so that Plumlee's data may be compared directly with data obtained in this research. The KR_{21} was calculated from data presented in her report, and all consistency and reliability coefficients adjusted downward to a test length of fifteen items. The means and variances were also adjusted to correspond to a fifteen-item test.

Table 2.3 Summary data from Plumlee [72] adjusted for test length

Test	Format	r	$\hat{r}$	$\bar{p}$	s	s^2	KR_{21}
W	SA	.686		.5236	2.87	8.25	.5856
W	MC	.640	.56	.5250	2.70	7.28	.5206
X	SA	.625		.4389	2.75	7.58	.5494
X	MC	.569	.48	.5528	2.72	7.42	.5367
Y	SA	.703		.5028	2.81	7.89	.5622
Y	MC	.517	.56	.5514	2.60	6.77	.4837
Z	SA	.640		.4319	2.75	7.56	.5494
Z	MC	.640	.48	.5861	2.70	7.28	.5359

II. A. 10. Summary of measurement theory

The probable error of a test score is some function of the standard deviation of a series of repeated measures. Since the same students cannot be repeatedly tested without learning, forgetting, practice, or fatigue affecting their test scores, and since the same

tests ought not be administered more than once to the same students, the probable error cannot be determined by the methods used in the chemical laboratory to calibrate an analytical instrument. The average standard error for a test given to a whole group of students can be estimated. It is a function of the length of the test, the spread in the distribution of scores, and the reliability of the test. The test length is not related to test quality. In contrast, test reliability and score distributions are intimately related to test quality characteristics such as difficulty level, variability in content, item discrimination, and guessing opportunity. Because reliability is related to variables which affect test quality while the standard error of measurement is not, investigation and quantification of measurement error was done through estimating and analyzing test reliability.

The most widely accepted operational definition of test reliability is the product-moment correlation between parallel forms. An internal consistency coefficient such as the KR_{20} is an adequate estimate of test reliability only when items within the test are as similar to each other as they are (or would be) to items on a possible parallel test. Since the tests given in CEM 130/131 cover a wide range of topics, their internal consistency is lower than their reliability. For these tests, therefore, internal consistency is not appropriate as a statistic with which to estimate errors in measurement. The correlation coefficient between two equivalent tests given to the same group of students was used to estimate test reliability.

The classical theoretical definition of test reliability is the ratio of true-score variance to total observed variance. Since random errors are uncorrelated with true scores, this definition may be restated: test reliability is equal to one minus the ratio of error variance to observed variance. With this alternate form of the variance-ratio definition of test reliability, the magnitude of the error variance and the effect on this error of test characteristics such as guessing opportunity were calculated from the parallel-forms reliability coefficients.

In the next section are set forth some of the techniques used to prepare test questions for the experimental comparisons presented in the results.

II. B. Writing and restructuring test items

The test questions which appeared on some of the tests which were compared in these studies did not require any changes in item format or number of answer choices. Other comparison schemes involving short-answer and multiple-choice questions with more than five response options often required substantial reworking of many test items. The conversion of multiple-choice questions into short-answer questions of a fully equivalent nature is quite easy for some items and difficult if not impossible for others. The expansion of the set of response options for a multiple-choice item is a trivial exercise for numerical problems and seldom truly difficult for other items. The creation of a response set for a short-answer item so that is may appear as a multiple-choice item on a test is also relatively easy although the multiple choice item thus formed may be too easy.

II. B. 1. Increasing the number of response options

There are two basic strategies used to create or expand the response set of a numerical problem. The simplest method is to provide a collection of numbers among which the correct answer can be found. No attempt is made to provide distractors representing logical though incorrect calculations. The more sophisticated strategem is to provide distractors which are the result of possible mistakes in algebra, arithmetic, or substitution. If more distractors are needed than are produced by this 'probable errors' approach,

sufficient 'random' numbers are selected to balance the set of answer choices. This probable errors method should be employed whenever the development of the response set is tractable to a logical plan.

These same strategies are also appropriate when an item is not a numerical problem. The following example demonstrate the application of the 'random numbers' and 'probable errors' methods for nonnumerical questions.

Example 1:

Which of the following is true when electricity is flowing in an electrolytic cell?

a. Anions migrate toward the cathode.
b. Anions migrate toward the anode.
c. Cations migrate toward the anode.
d. Electrons, not ions, move through the solution.

The expanded set of possible choices is:

a. Anions migrate toward the cathode.
*b. Anions migrate toward the positive electrode.
*c. Anions migrate toward the anode.
d. Anions migrate toward the negative electrode.
*e. Cations migrate toward the cathode.
f. Cations migrate toward the positive electrode.
g. Cations migrate toward the anode.
*h. Cations migrate toward the negative electrode.
i. Negative ions migrate toward the cathode.
*j. Negative ions migrate toward the positive electrode.
*k. Negative ions migrate toward the anode.
l. Negative ions migrate toward the negative electrode.
*m. Positive ions migrate toward the cathode.
n. Positive ions migrate toward the positive electrode.
o. Positive ions migrate toward the anode.
*p. Positive ions migrate toward the negative electrode.
q. Electrons migrate toward the cathode.
r. Electrons migrate toward the positive electrode.
s. Electrons migrate toward the anode.
t. Electrons migrate toward the negative electrode.

(*) denotes correct response

The above twenty possible responses include eight correct answers and twelve distractors. This selection is obviously not exhaustive. The response set for the above item is subsceptible of expansion by inclusion of various logical permutations of the related concepts. Even for a question in an area not narrowly restricted to a small set of related concepts, the method of probable errors may be employed to construct the answer set.

Example 2:

Which of the salts listed below could be used to produce an acidic solution?

1. K_2HPO_4 2. NH_4Cl 3. LiHS 4. KF

The expanded set of possible answers is:

a.	only 1	i.	2 and 4
b.	only 2	j.	3 and 4
c.	only 3	k.	1, 2, and 3
d.	only 4	l.	1, 2, and 4
e.	1 and 2	m.	1, 3, and 4
f.	1 and 3	n.	2, 3, and 4
g.	1 and 4	o.	all of these
h.	2 and 3	p.	none of these

From these distractors, nine may be selected to convert a four-choice question into a ten-choice question. This method of increasing the number of answer choices does not apply when the question is worded such that the answer is perforce unique.

Example 3:

Which of the following gases has the lowest critical temperature? The addition of possible distractors to this item is not strictly planned according to what probable errors the student might make although

that would influence the selection of choices. Rather, distractors are chosen randomly and included in the answer set as long as the correct answer remains unambiguously right and as long as a distractor is not so obviously wrong as to be ineffective.

Example 4:

> A pure substance which melts over a wide temperature range is likely to be
>
> a. crystalline
> b. molecular
> c. amorphous
>
> An expanded set of responses:
>
> a. hexagonal closest packed
> b. an ionic crystal
> c. of small formula weight
> d. symmetrical
> e. of very low density
> f. a molecular crystal
> g. cubic closest packed
> h. amorphous
> i. hydrogen bonded
> j. an organic compound

Though many of the distractors refer to substances which might melt over a wide range, only a substance which is amorphous always behaves in this manner. When the supply of logical and directly related distractors is exhausted, profitable use may be made of such essentially irrelevant but dimly related distractors as illustrated in Example 4.

II. B. 2. Interconversion of multiple-choice and short-answer items

Numerical problems are readily converted between the multiple-choice and short-answer formats. Seldom is even a minimal change

needed in the wording of a problem. A multiple-choice question is changed into a short-answer question by elimination of the answer set. A short-answer item is transformed into a multiple-choice item by construction of a response set utilizing the methods just discussed. The conversion of nonnumerical multiple-choice questions into short-answer questions frequently requires changes in the wording of the question. Often the answer choices which are provided by a multiple-choice question limit the scope of a question which would otherwise possess myriad correct responses. Sometimes the responses include information necessary to the solution of the problem. Two examples are presented here: Example 5 is easily converted to short-answer format whereas Example 6 is converted only after modification substantial enough to radically alter the character of the item.

Example 5:

Which of the following is a detailed ionic equation which represents the reaction of soluble $CrCl_3$ with the hydroxide ion to form insoluble chromium (III) hydroxide?

a. $Cr^{3+} + 3Na^+ + 3Cl^- + 3OH^- = Cr(OH)_2 + 3Na^+ + 3Cl^- + OH^-$

b. $3OH^- + Cr^{3+} = Cr(OH)_3$

c. $Cr^{3+} + 3Cl^- + 3Na^+ + 3OH^- = Cr(OH)_3 + 3Na^+ + 3Cl^-$

d. $CrCl_3 + 2NaOH = Cr(OH)_2 + 2Na^+ + 3Cl^-$

e. $CrCl_3 + 3Na^+ + 3OH^- = Cr^{3+} + 3OH^- + 3NaCl$

Conversion of Example 5:

Write the detailed ionic equation which represents the reaction of soluble $CrCl_3$ with sodium hydroxide to form insoluble chromium (III) hydroxide.

Example 6:

Consider the following balanced equation:

$3H_2S + 2H^+ + 2NO_3^- = 2NO(g) + 4H_2O + 3S(s)$

a. It is a net ionic equation
b. It is a detailed ionic equation
c. It is a molecular equation
d. If $2Na^+$ is added to both sides, it will then be a detailed ionic equation
e. If $3H_2S$ is replaced by $6H^+ + 3S^{2-}$, it will then be a detailed ionic equation
f. If $2H^+$ and $2NO_3^-$ are deleted, it will then be a molecular equation

So simple a change as rewording the first phrase in the item stem will not transform this question into a short-answer format. The resulting short-answer may be confusing or cumbersome or both. Several attempts at converting Example 6 into short-answer format are offered here to illustrate this problem.

Conversion 1 of Example 6:

Consider the following balanced equation....
What type of equation is this? If this equation is not a true example of any type, what must be done to make it a detailed ionic equation?

Conversion 2 of Example 6:

Consider the following balanced equation....
State whether this equation is a detailed ionic equation. If it is not, what must be done to make it a detailed ionic equation?

Conversion 3 of Example 6:

Consider the following balanced equation....
Is this equation molecular, net ionic, or detailed ionic?
If it is none of these, what must be done to make the equation a detailed ionic equation?

The first conversion of Example 6 is, to say the least, ambiguous. Some of the ambiguity may be overcome by sacrificing some of the scope of the question as shown in the second conversion attempt. This short-answer question no longer includes references to molecular or net ionic equations. The third conversion contains almost all the elements of the original multiple-choice question, but is merely a multiple-choice question in short-answer clothing. Scoring ease has been sacrificed in this case to gain a small amount of ambiguity. When an item is as intractable to conversion as is Example 6, it is best not to attempt direct conversion but to select another question altogether.

The creation of a response set for a nonnumerical short-answer item is, in contrast, straightforward. On rare occassions the multiple-choice item thus created will be inappropriately easy. This stems from two related causes -- either the number of possible answers is too few or the clues provided by the answers are too strong.

Example 7:

The point of zero amplitude on a wave is called the...

A possible response set:

a. crest b. node c. trough

Example 8:

The smallest particle of light energy is called the...

A possible response set:

a. quantum b. electron c. gamma ray

Multiple-choice questions constructed by simple conversion of short-answer questions with such restricted answer domains will seldom be comparable in difficulty. Also, rewording the item to make it more than trivial will also make it less than equivalent.

II. B. 3. Limits of structural change

Multiple-choice questions call for the selection of an answer from among a set of possible choices according to a problem or idea stated in the item stem. Whether the item can be converted into short-answer format depends on the nature of the response set. Three general classes of response sets are: (1) a restricted set with a clearly defined descriptor, (2) a well-defined but essentially unlimited set, and (3) an arbitrary or undefined set of responses. Several examples of these follow.

Example 9:

The cause of the boiling point elevation of a solution is

a. vapor pressure increase
b. osmotic pressure increase
c. freezing point depression
d. thermal conductivity
e. vapor pressure decrease

Example 10:

Dissolving potassium nitrate in water is an endothermic reaction. The amount of KNO_3 which dissolves could be increased by

a. cooling the KNO_3 before adding it to the water
b. heating the water while adding the KNO_3
c. adding a large excess of KNO_3 to the water
d. vigorously stirring the water while adding KNO_3

For these two examples there are class discriptors or prompts available which direct the student to consider only the restricted set of possible responses.

Conversion of Example 9:

The colligative property which is the direct cause of the boiling point elevation of a solution is...

Conversion of Example 10:

Dissolving potassium nitrate in water is an endothermic reaction. According to LeChatelier's principle, the amount of KNO_3 which dissolves could be increased by...

The underscored prompts remove the ambiguity which might be present in their absence. If the response set suggested by the prompt is too small, however, the item may be of much less value. The conversion of Example 1 on p. 80 from multiple-choice format into short-answer format illustrates this 'item trivialization' effect.

Conversion of Example 1:

When electricity is flowing in an electrolytic solution, the anions migrate toward which electrode?

There is no way to remove ambiguity without making the converted form considerably easier than the original multiple-choice item. Alternately, when the number of possible responses is too large, an attempt to restrict the possible choices by qualifying the short-answer question produces a like effect. Not restricting the set so as to avoid trivialization may lead to conversion of a multiple-choice item into a free-response item which is better characterized as essay rather than as short-answer.

Example 11:

In concentrated ionic solutions,

a. ions move faster than in dilute solutions
b. ions are more restricted in their motions due to interionic attractive forces
c. solute molecules are never completely ionized
d. ions affect colligative properties to the greatest extent

Conversion of Example 11:

Discuss the colligative properties of electrolytes and nonelectrolytes as a function of concentration.

The conversion of Example 11 into a short-answer question founders because there is no narrow definition of the several possible responses in the multiple-choice item.

Example 12:

The results of simple tests made upon several solutions are listed below. Which result indicates that the solution was initially saturated?

a. The solution was warmed and solute crystallized out
b. Pure solute was added and some of it dissolved
c. A crystal of solute caused crystallization to occur
d. Solute was added and no change in concentration observed

Conversion of Example 12:

Suggest simple experimental tests that would allow you to distinguish among solutions which might be unsaturated, saturated, or supersaturated.

The conversion of Example 12 again leads to an essay question rather than a short-answer question.

When the domain of answer choices is very large, the selection of choices for a response set is essentially arbitrary. The response set is also arbitrary when the nub of the question is a comparison,

since the selection of responses to be compared is discretionary. Such multiple-choice questions as Examples 13 and 14 cannot be converted into parallel short-answer questions.

Example 13:

In the following list, which compound is a weak electrolyte?

a. NaCl b. NH_4OH c. $LiClO_4$ d. HI

Conversion of Example 13:

Write the formula of a weak electrolyte.

Example 14:

Which of the following ions should be most strongly solvated in aqueous solution?

a. F^- b. S^{2-} c. Br^- d. O^{2-}

Conversion of Example 14:

Write the formula of an ion strongly solvated in aqueous solution.

These possible short-answer questions certainly are not parallel to their multiple-choice analogs. These last two examples of multiple-choice questions are in fact testing for achievement of a skill which cannot be explicitly tested in any other way. The ability to make a selection from among several options is pertinent to many practical situations. Which indicator to use in a titration, what reagents to combine for a buffer solution, what salt to choose for an electrolytic bath, which experimental conditions will produce a desired effect -- these are only four examples of innumerable applications of the ability to select.

II. C. Experimental design

II. C. 1. Introduction

Testing procedures in CEM 130/131 require large numbers of examinations to be generated, administered, scored, and recorded. To facilitate this task, examination generation and record keeping are done by computer. Scoring is usually done by hand because scoring machines are not readily available. Exams are administered to large groups of students in lecture halls and are scored immediately thereafter.

The examinations used in these courses consist almost entirely of multiple-choice questions. However, it is believed that the typical multiple-choice items (and true-false items in particular) permit guessing to such an extent that fifteen-item exams are too easy and too unreliable. The purposes of these reliability studies are to investigate the reliabilities of various types of exam questions under different conditions, to compare reliabilities of short-answer on the one hand and multiple-choice items with different numbers of choices on the other, and to compare means on exams with different question format distributions.

II. C. 2. Selection of item format

Three basic item formats appear on a typical chemistry exam -- multiple-choice, short-answer, and problem or essay. Each of these types has drawbacks and strong points, and for each type there are

topics which lend themselves with ease to a particular format; for other areas it is very difficult to write a good item in that same format. Multiple-choice and short-answer items are both classified as objective items* because of the straightforward manner in which they are scored. The answer to an objective item is -- or should be -- clearly right or wrong, and no partial credit is given. This is in contrast to the problem or essay question where the grader evaluates a written response subjectively or gives credit in a problem based on method or partial solution.

In CEM 130/131 the student is given the opportunity to take tests once a day during an approximately ten day period with his grade on the test determined by a preset grading scale. The student may repeat an exam until satisfied with his score, or until the closing date. Only the last score (not the highest score) is the one on which his grade is based. During a typical exam week one to four thousand tests may be administered, scored, and entered into student records. Since such large numbers of tests are used, all items are of the objective format so that they can be rapidly scored by nonexpert graders.

*The true-false item may be treated for computational purposes as a special case of the multiple-choice item with only two options. A matching item is a complex set of interlocking multiple-choice questions with decreasing numbers of options per item. Questions with more than one correct answer can be considered a collection of true-false questions that may contain more than one true statement. Matching and multiple-answer items cannot be analyzed because of the interlocking nature of their responses. None of the items used in CEM 130/131 or in these studies are of the latter two types, and very few are true-false items.

II. C. 3. Components of observed test variance

The observed test variance may be distributed into several categories which are uncorrelated. The components listed in Table 2.1 and restated below are of varying magnitude and importance.

Table 2.1 Components of observed test variance

Symbol	Source
s^2_T	differences in knowledge (T = true score)
s^2_{fl}	learning and forgetting (fl = fluctuation in true score)
s^2_g	guessing variations among students (g = guessing)
s^2_{adm}	effects of day, room, etc. (adm = administration)
s^2_{eq}	unrepresentative coverage of topics (eq = equivalence)
s^2_{sc}	subjective scoring errors (sc = scoring)
s^2_m	memory of previous test questions (m = memory)
s^2_R	all remaining sources of error (R = residual)

Error variance attributable to scoring and to administrative conditions will approach zero under the conditions prevailing in this system and will not be further considered.

Variance caused by fluctuation in true scores is also assumed to be negligible when the interval between tests is as short as it is for these reliability studies. The mean time between all items on the two tests is never more than one hour and for most examinees it

averages about forty minutes.* Fluctuation in true score is not classed as error since a perfectly precise test would accurately register a difference if the score truly fluctuated. However, any fluctuation is likely to be diagnosed as error produced by a lack of equivalence between tests. There is no way to determine whether a change in observed score is caused by a change in what the student knows or by a change in what the test measures. Since true-score fluctuation cannot be separated from test-content fluctuation, it will not be considered explicitly as an important source of variance different from variance caused by a lack of equivalence between tests.

The effect of memory of responses given previously is of unmistakable consequence when the same items appear on a subsequent test. But there is also a gray area between the error variance of memory and the lack of validity of the items. Memory is one extreme; the student encounters an item a second time and remembers how he answered it the first time. Knowledge is at the other extreme; the student encounters a 'different' item testing for the same knowledge and is again able to solve the problem. If however the student is able to answer a question based not on true knowledge but on some superficial item characteristic, whether this leads to memory variance or true-score variance depends on what one classifies as superficial.

*Only one of the correlation studies used more than two tests; on that single occasion three tests were administered in three different orders to three groups of students. The first and third tests would then have a mean separation time twice as long as consecutively taken tests.

Certainly when a student is able to answer a question due to testwiseness rather than knowledge of chemistry, the 'knowledge' he demonstrates is superficial. Yet consider this example:

> Examic acid is a monoprotic acid with $K_a = 1.00 \times 10^{-6}$. What is the pH of a solution which is one molar in both examic acid and potassium examate?
>
> (1) 1.0 (2) 6.0 (3) 14 (4) 9.9 (5) 8.2

A student could arrive at the correct answer in many ways without making a random guess. The method which the item writer probably had in mind is to set up the equilibrium expression, solve for the concentration of hydrogen ion, and convert to pH units.

Another method is based on a restricted principle. When a solution consists of equimolar quantities of a weak acid and its salt, the pH of the resultant buffer is equal to the negative logarithm of the ionization constant of the acid. While this second method tests a slightly less sophisticated kind of knowledge, the instructor would likely accept this measurement as valid and meaningful.

Still another method is based on a less rigorous chain of logic. A solution of a weak acid and its salt is a weakly acidic buffer. The scale of pH is defined such that a solution with a pH of seven is neutral and smaller numbers denote acidic solutions. The second answer choice is the only pH value which is weakly acidic. This third method of solution is evidence of a level of knowledge much lower than the item is designed to test. The knowledge is not trivial, but it is not as

advanced as intended. This last method is in the gray area; it borders on 'chemistrywiseness' rather than 'testwiseness'. Even if the student is only able to narrow his options to the two values which are acidic pH and guess which is correct, does not this show some familiarity with acids, buffers, and the pH scale? Perhaps even such intuitive acquaintance with the subject matter is a valid educational goal.

Because the definition of memory variance for other than retesting with identical items leads into such a gray area, the consideration of memory variance will be restricted to the test-retest situation. For all reliability studies where nonidentical tests are compared, error variance due to memory is assumed to be zero.

Thus far, four of the variance components from Table 2.1 have been eliminated from general consideration: scoring, administration, fluctuation, and memory. Four categories remain to be considered: true score, guessing, equivalence, and residual. True-score variance will equal zero only when all students know the same amount of chemistry.* Residual variance is that category containing all sources of variance not considered explicitly; it is a theoretical bet-hedge so as to avoid the accusation "But you forgot..." Residual is also the convenient residence for all sources of error other than the single category being investigated. For example, when guessing variance is

*Under a mastery system of learning, most students achieve the same level of knowledge -- mastery -- and variation among true scores is perforce low or zero. The estimation and interpretation of reliability and measurement error for such tests is more difficult.

being studied, residual variance will include the variance caused by nonequivalence -- certainly not an inconsequential source. Guessing and equivalence are the two major sources of errors in measurement which are investigated in this research. Their importance and the method of their estimation are discussed in the next section.

II. C. 4. Experimental design

Errors caused by guessing ir by nonequivalence are important for two reasons. First, since students are allowed to retake tests on which they perform poorly, the test score as a reliable measure of knowledge and not the lucky guess is important. Second, since different forms of an exam are taken by different groups of students -- yet graded on the same scale -- all parallel forms should be equal in difficulty and representative in content. Both of these reasons concern the stability of the measurement: the stability of the individual score, and the stability of the test mean. Only when a test is reliable in both senses will a test score be precise and accurate on an absolute rather than a relative scale of achievement.

The test reliability and the measurement error resulting from guessing and nonequivalence are estimated from internal consistency coefficients, parallel forms correlations, and summary statistics for each test administration.

II. C. 5. Comparison schemes

When tests are compared, there should be item-by-item content correspondence between the tests, and all items should be the same format. This allows the purest estimates to be made of the contrast between two formats or the reliability of a single format. At the same time, the tests used and the experimental setting should be identical to that typically found in the classroom. This allows the results of the research to be generalized to the actual testing situations. As is often the case, conditions which make experiments more theoretically rigorous also make the results less readily applicable to the classroom, and experimental designs which mirror the classroom may contain too many uncontrolled variables and complex interactions which make it difficult to draw sound conclusions.

The tests used in this research are not different in content from those regularly used in the course. Often, however, extensive changes were made in the structure and format of questions (according to the procedures outlined in Section II. B). Because some items could not be restructured in keeping with the format for a particular comparison, seldom are the fifteen-item tests used in these studies of one format. Rather than manipulate test content artificially, some structural design purity was surrendered. This approach sacrifices some mathematical rigor and computational ease in order that the results be confidently applicable to the course environment. Statistics were calculated when appropriate for subtests of items more nearly identical

in format. The data from the thirty-two comparisons which were made are grouped into five general categories.

II. C. 5. a. Comparisons of parallel tests of five-option multiple-choice items

This first category includes nine structurally unchanged tests consisting almost entirely of five-choice items. There are fifteen comparisons in this category, some of which differ only in the order in which the two tests were administered. The correlation coefficient between two tests of this type is an estimate of the reliability of the five-choice item. The estimate will be less than 1.00 to the extent that nonequivalence and guessing contribute to error variance. It should also be noted that because a fifteen-item exam has such a restricted scale, the maximum possible correlation will be less than unity unless test means and standard deviations are equal.

II. C. 5. b. Comparison of parallel tests with more than five options per item

The tests in this category had the number of answer choices increased to between six and ten from the original four or five. The six comparisons are among seven different tests. As in the first scheme, error variance is subdivided into primarily guessing error and nonequivalence between forms. The correlation coefficient between two tests in this category will show the effect of increasing the number of response options on test reliability.

II. C. 5. c. Comparison of parallel tests of short-answer items

The measured reliability of the short-answer format is the maximum possible value for test reliability when the only variable allowed to change is the number of response options. In this comparison, it is assumed that errors due to guessing are zero, and any observed error variance is therefore essentially that of nonequivalence. The four comparisons in this category are among four tests which consist predominantly of short-answer items. These items were prepared from the original multiple-choice format with as little change in wording as possible.

II. C. 5. d. Comparison of short-answer items with parallel multiple-choice items

The comparisons in this category are between short-answer items and multiple-choice items with eight and ten options. The guessing error introduced by the multiple-choice format will be only half what it would be if both tests were multiple-choice. The harmonic mean of the number of choices when two ten-option tests are compared is ten; the harmonic mean when a ten-option test and a short-answer test are compared is twenty. This is a way of indirectly doubling the mean number of answer choices so that the reliability of more-than-ten response options can be estimated without having to construct such items.

II. C. 5. e. A variation of a test-retest comparison

The four comparisons among six tests in this category are a variant of the test-retest method. When a student encounters the same test question a second time, he may give the same answer not because he knows the answer both times but rather because he remembers the answer he gave the first time. A short-answer question does not provide answer choices to be remembered, so that the only aspect of memory which applies to a second test is whether the student can still remember how to answer a question he could answer on the first test.

Each group of students in these comparisons was administered two tests. The first test was a short-answer test created by deleting the response options from multiple-choice questions, and the second test was the identical set of items 'restored' to multiple-choice format. Inasmuch as the content of the two tests is unchanged, there should be no error variance due to nonequivalence of content.*

These comparisons will not be a valid test of how much easier the multiple-choice format is in comparison with the short-answer format because each student will have had more time to work on the multiple-choice questions.

*There may be some arcane nonequivalence of content introduced by the change from a recal to a recognition item. The use of the test-retest method almost always involves some such troublesome idiosyncratic kind of error variance. This is why Cureton [73] said, "So far as I can tell, the test-retest coefficient has no clear interpretation under the weak true score [classical] theory."

II. C. 6. Administrative procedure for 'Special Option' exams

Student participation in these experiments was voluntary. Notices of the date, time, and place of a 'Special Option' exam were prominently posted during the week prior to each occurrence. To encourage student participation, students were informed that only the higher of the two test scores would be entered into their records.

As students filed into the testing room, they were fiven the first of two exams. When a student finished the first exam, he turned in his answer sheet and was given the second exam. When he finished the second exam, the student returned his second answer sheet and left the testing room.

When two forms were administered concurrently, they were distributed to the students with random irregularity; true random assignment would have been impossibly cumbersome.*

*Both forms were distributed simultaneously to students as they entered. On some occasions exam forms were randomly collated in the same stacks, but on most occasions two or more proctors, each with a separate test form, handed out exams. Never was one form distributed first followed by the second form, nor was any conscious discrimination by sex practiced.

II.D. Results and discussion

The ideal method of determining the reliability of a chemistry test would be to administer several flawless examples of specific tests to a perfect group of students. This group would be perfect in two ways: knowledge of chemistry would be represented at all levels from no knowledge of the topics tested to all that should be known about them; further, the extent of each student's knowledge of chemistry would be absolutely constant during testing and the students would be completely unchanged by the testing experience. Such groups of students and such collections of tests do not exist. The ways in which the real departs from the ideal introduce several different sources of error into the measurement process.

Although one may construct several different forms of a chemistry test, they are all not necessarily equivalent. Any differences between them in the emphasis of topics covered on each exam introduce error variance due to this nonequivalence. Questions which appear on a test can be thought of as chosen from a universe of all possible questions about the entire scope of chemical topics covered by the examination. Only if all possible test items are identical in content and difficulty -- which they obviously are not -- would any seires of samples drawn from this population be perfectly equivalent. The internal consistency is a measure of the possible sampling error which might exist because all items are not identical.

A second kind of sampling error occurs when the group of students tested to establish reliability is not representative of the population of students that could have taken the exam. The measurements based on unrepresentative samples may not correspond to population parameters. Consequently, the test means and variances or item difficulties and discriminations for a small atypical subgroup are likely to differ from those of the total course enrollment.

It is also unlikely that the student is immutable in his knowledge of chemistry. He studies and learns more, and when active study ceases, he begins to forget. Often the questions on a test provide the student with new insights into his knowledge; that is, the act of measurement may itself change what is measured. This fluctuation in a student's knowledge is reflected by a real change in his 'true score.' Nevertheless, there is no prima facie way to distinguish a true change in what the student knows from either a change in test content or from the random error of an imperfect test.

The contributions of these variables to error in measurement and their effects on test reliability have been mentioned in the theoretical discussion (Section II. A). The effects of position, form, and group when more than one test is given to a group of students, as well as the effects of variables such as the number of answer choices and test content are examined in detail in the following pages.

II.D.1. Effects of position, form, and group

II.D.1.a. Introduction

The correlation between two tests can be influenced by factors other than the reliabilities of the tests. Among these are the effect of the order of test administration (position effect), the differences in content between forms (forms effect), and the differences between groups in ability and range of abilities (group effect).

Position effect. The order in which tests are taken and the time between their taking may affect test reliability because of fatigue, practice, or forgetting. These effects are estimated by the differences in test means and intercorrelations between tests administrated in different positions, either successively or separated by a third test. Tables of these results are listed in Appendix B along with estimates of the interaction between position and form or group.

Forms effect. The correlation between tests may also be decreased to the extent that different forms of an exam ask questions about different topics. The forms effect is therefore the direct measure of equivalency, and test equivalence is a fundamental assumption for reliability estimation by parallel forms correlation. Content equivalency is not demonstrated merely when tests have the same mean -- nor disproved when means are significantly different. A yardstick and a meterstick provide quite different raw estimates of length though both are measuring the same 'topic'. Equivalence of difficulty (as opposed to content) is discussed in Section II.D.4.

Group effect. A third important factor which influences test reliability is the character of the group of students tested. A heterogeneous group will have a large true-score variance since there are large real differences in ability among students. The larger the true-score variance, the likelier the reliability will be high since reliability is the ratio of true to observed variance. When two groups are tested, the test given to one group may have a higher reliability simply because that group contained a greater spread of abilities. Comparing the same form of a test across different groups averages much of this difference. This group effect is estimated by the difference in test variances between groups which took the same test. These effects are summarized in Appendix B.

II.D.1.b. Observed test reliabilities

The reliability coefficients and confidence intervals for various test administrations are displayed in Figure 2.1. The confidence interval extends two standard deviations above and below each coefficient; this interval encompasses a 95.44 percent probability. The width of the confidence interval depends only on the size of the group tested. Administration Code A has been divided into the three groups (W, X, and Y) which took the tests together. These data are taken from Appendix B, Tables B.3 and B.12.

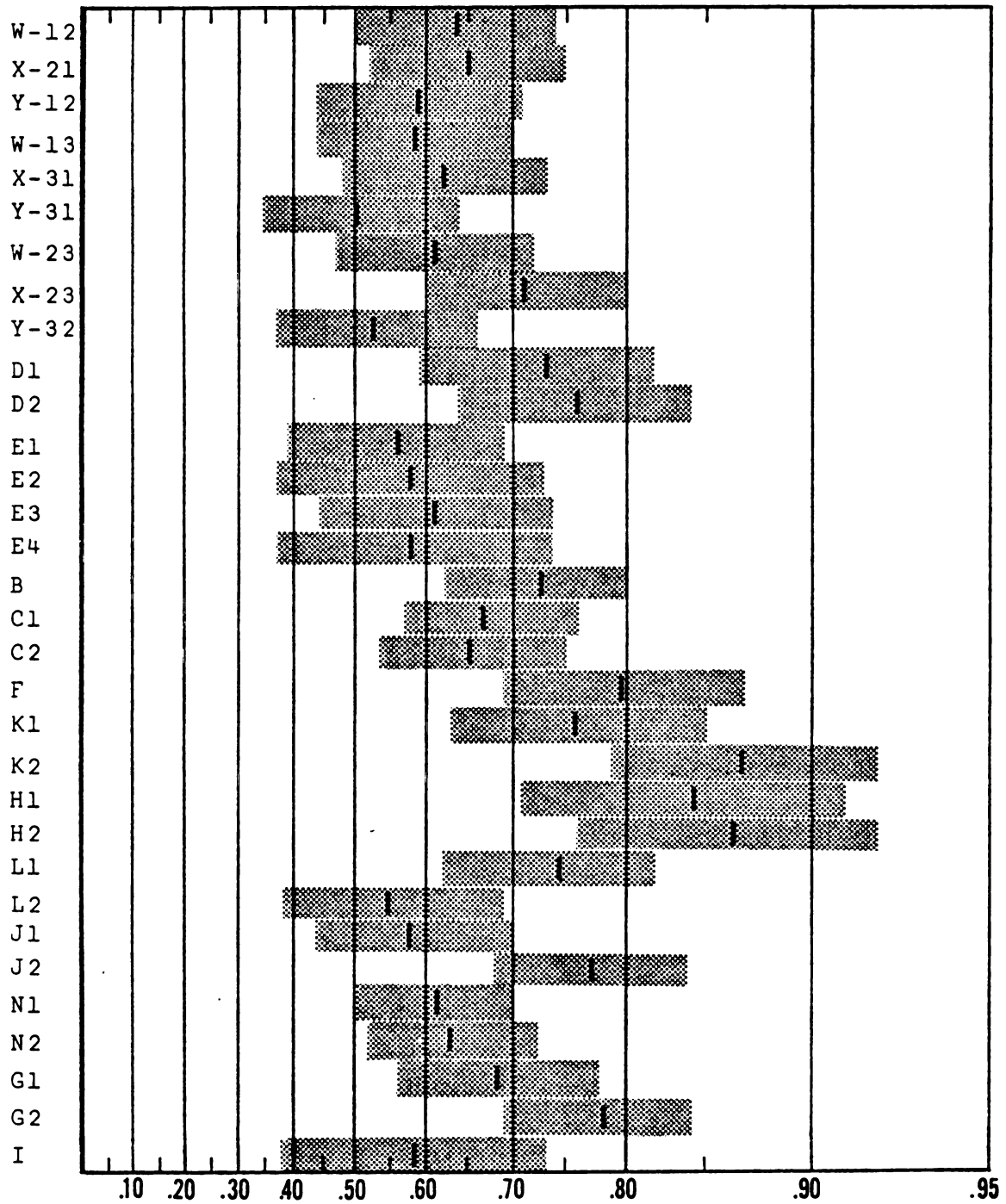


Figure 2.1 Reliabilities and $\pm 2\sigma$ confidence intervals for tests of fifteen items listed by test code (see Appendix B).

II.D.1.c.i. Discussion of non-zero effects: Position

The analysis of variance (ANOVA) technique is inapplicable for these data because each of the 'cells' in the design is not independent. The same group is administered different forms of an exam so that reliability may be estimated by correlation. If the exams were correlated 1.00, then any difference in test means would be statistically significant. The difference cannot be due to differences in the groups (the same group takes both tests) and cannot be due to random error variance on one exam which differs from random error on the other since they intercorrelate perfectly. However, if the intercorrelation among cells is less than unity, the size of the correlation coefficient is inversely related to the significance of a difference in test means. The threshold for the significance of a difference is always less than that calculated with the assumption that the two means are not correlated. Thus the analysis of variance will underestimate any differences present. In contrast, multiple t-testing will overestimate any differences because when many t-tests are done on a set of data, some results will be significant merely by chance. Therefore, only a strong pattern of significant results will be interpreted as indicating a significant difference; single significant results will be lightly taken.

Although inappropriate because differences are underestimated, the analysis of variance is listed in Table B.13 in Appendix B. Even though the effect is underestimated, it is unlikely that the position effect could inflate enough under another technique to become significant.

The position effect is also nonexistant when the more liberal multiple t-testing procedure is used. For 24 comparisons across 19 groups and between 18 separate forms, not a single significant difference is found. When groups are pooled for each administration of the same forms in different orders, the eleven differences contain only one which is significant and this is at an alpha almost greater than .05 -- i.e. bordering on nonsignificance. These results are tabulated in Appendix B, Tables B.7, B.8, and B.14.

II.D.1.c.ii. Discussion of non-zero effects: Form

As was the case for the position effect, multiple t-tests were used to discern any differences due to the lack of equivalence of purportedly equivalent exams. Because the same group was given each form, the means are not independent and an adjusted standard error for correlated or dependent samples was again used. The differences and significance tests for the forms effect are listed in Tables B.4, B.5, and B.15 in Appendix B. Each of the differences is between two forms of an exam given to the same group. Fifteen of the eighteen differences are significant. Thus, even expecting a significant result occasionally by virtue of the multiple t-test procedure, it is clear that there is a forms effect -- the pattern is too strong to attribute to chance.

II.D.1.c.iii. Discussion of non-zero effects: Group

There are two possible group effects: overall level of ability, and range of ability. There are also two ways these possible differences can affect reliability estimates. The total group of students at any single test administration may be unrepresentative of the general student population by being more able or heterogeneous or less so. The distribution of alternate forms of an exam may divide the group of students into dissimilar subgroups. In statistical terms these two areas of concern are selection (from the population) and assignment (to the treatments), and both should be random.

There is no method by which these data may be analyzed which will show whether or not the actual groups tested are representative. Whether the subgroups of each administrative code are equivalent is determined by testing differences between means summed for each division of a group across all tests taken by the students in the group. These differences and tests of significance are presented in Tables B.6, B.9, and B.16. Of the eleven differences, the only one which is significant is between two groups which took two exams, only one of which was the same for both groups. Here, the group effect is compounded with the forms effect. Thus it appears evident that there is no group effect in the sense of systematic important differences in 'assignment' to each order of testing. Whether the method of self-selection of students for these special option exams produces extremely nonrepresentative groups is an unanswered question.

The sizes of the groups tested are for the most part 'large', and the ranges of scores received on the examinations are not unusually restricted or dispersed. The nature of the effects discussed in these pages is primarily relative -- that is, whether a particular test characteristic increases reliability or test mean compared with another characteristic. It would seem reasonable that differences in group heterogeneity or average ability might affect absolute values for the various effects but not the direction of a relationship. Regardless of how poorly a thermometer is calibrated, if it functions at all, one can easily determine which is the hotter of two objects.

II.D.1.d. Summary of position, form, and group effects

There is clearly no position effect and -- just as clearly -- there is a forms effect. It also seems reasonable to conclude that there is no significant group effect when the groups are large. There are slight interactions between form and group, one group doing slightly better on a particular form than another group, but this interaction effect is not often seen and is never as large as the forms effect caused by differences in content between alternate forms of an exam.

II.D.2. Effects of some examination characteristics on reliability

The estimated test reliability may relate in a complex manner to other examination characteristics such as chance for guessing, mean difficulty, or spread of abilities. Eight variables including these three

were used to predict the correlation coefficient through a multiple regression computer program. The summary of the output from the program is listed in Table 2.4. The standardized Z-score beta weights indicate the separate importance of each variable in the prediction equation. The four best predictors are the standard deviation, the number of multiple-choice answer options, the intercorrelation among the test items (content homogeneity), and the difference in chance for guessing between two tests correlated. The standard deviation and item intercorrelations are characteristics of the test and group, while the guessing possibility and differences in guessing possibility between tests are structural characteristics independent of the group tested.

The entire eight-variable prediction equation accounts for sixty-two percent of the variance in correlation coefficients; the ramainder (thirty-eight percent) is unaccounted for by these variables* (and perhaps by any variables). The opportunity to guess -- as indicated by the multiple-choice-equivalent**(MCE) number of response options -- accounted for the most variance (34 percent) of any single variable. This effect of the number of response options will be discussed in greater detail in Section II.D.3.c.ii.

The test mean did not contribute much to the prediction equation, although it did correlate negatively with reliability (- 0.43). The expected relationship between test mean and reliability is that as the

*Variation in the degree of parallelism between equivalent tests is the likeliest cause of the remaining variance in coefficients.

**[harmonic mean]

mean increases above the optimal mean (halfway between chance and perfect), the reliability increases because there is less guessing. There are not enough values for means above and below expected optimal means to test systematically this theoretical conclusion. The mean as an independent variable was entered into the regression equation to guard against any effect peculiar to these specific data. There was none.

Another variable which might have had a sizeable effect on estimated reliability is the percent of the test items which are problems, as opposed to nonnumerical questions. This dichotomization of items into problems and nonproblems is a simplification of the factor structure. As mentioned by Cronbach earlier, tests with a single strong factor have highest reliabilities, while tests with complex factor structures, especially tests of stratified content, have lower reliabilities unless the tests are long. Mathematical calculation is a factor which all problems have in common; thus problems in different chemical topics are somewhat similar by virtue of their underlying mathematical nature. The strength of this math factor may be great enough so that distinguishing between questions which are and which are not problems might be of predictive value. On the other hand, sets of items which are not problems may have a complex assortment of weaker factors. The strength of the general factor would depend on only how similar in content the questions are, and not whether they are problems.

The effect of simple versus complex factor structure is not evident in this treatment of correlation coefficients by regression analysis. Its subtle nature is more apparent, however, when the reliability of subsets of items are discussed in Section II.D.3.

The most important general result of this regression analysis is the demonstration that the number of answer choices determines to a great extent how reliable a test will be. The subdivision of variation in test reliabilities into its major fractions is visually displayed in Figure 2.2.

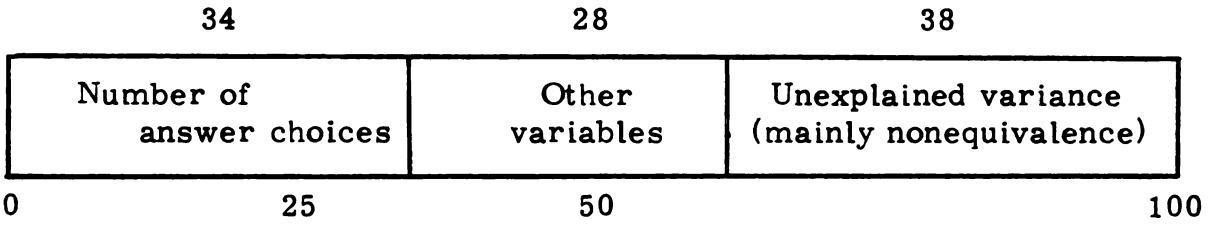


Figure 2.2 Percentage distribution of sources of variation in reliability coefficients from linear regression of eight variables and parallel forms correlation coefficients.

Table 2.4 SPSS Regression output summary

VARIABLE	SYMBOL	β-WEIGHT	Z_β	VARIABLE	SYMBOL	β-WEIGHT	Z_β
STDDEV	s	.0231	.5148	KR20	KR_{20}	.1381	.1100
MCEQUIV	MCE	.0050	.5043	DELTMEAN	$\Delta\bar{p}$	-.2213	-.0934
INTRCORR	$\bar{r}_{ij}$	-.6351	-.3525	PCTPRBL	100xfp	-.0002	-.0453
DELTMCEQ	ΔMCE	.0047	.1941	MEAN	$\bar{p}$	-.0189	-.0172

INTERCORRELATION MATRIX

	STDDEV	MCEQUIV	INTRCORR	DELTMCEQ	KR20	DELTMEAN	PCTPRBL	MEAN
STDDEV	1.0000	.5616	.8244	.0818	.9192	-.0923	.4256	-.5047
MCEQUIV	.5616	1.0000	.4480	.2088	.4524	.1515	.2849	-.2297
INTRCORR	.8244	.4480	1.0000	-.0317	.8774	-.1455	.4926	-.1191
DELTMCEQ	.0818	.2088	-.0317	1.0000	.0271	.5201	-.1277	-.3704
KR20	.9192	.4524	.8774	.0271	1.0000	-.1074	.4443	-.3318
DELTMEAN	-.0903	.1515	-.1455	.5201	-.1074	1.0000	-.0078	.0866
PCTPRBL	.4256	.2849	.4926	-.1277	.4443	-.0078	1.0000	-.0254
MEAN	-.5047	-.2297	-.1191	-.3704	-.3318	.0866	-.0254	1.0000

SUMMARY STATISTICS

STEP	VARIABLE ENTERED	F TO ENTER OR REMOVE	α	MULTIPLE R	R SQUARED	R^2 CHANGE	SIMPLE R
1	PCTPRBL	.1657	.686	.1683	.0283	.0283	.1683
	DELTMEAN	.5536	.461	.1864	.0347	.0064	.0788
	MEAN	.0109	.918	.4744	.2250	.1903	-.4318
	MCEQUIV	16.3883	.000	.7492	.5613	.3363	.6947
	INTRCORR	2.0816	.156	.7545	.5693	.0080	.3774
	DELTMCEQ	2.0528	.159	.7585	.5734	.0061	.3065
	KR20	.1416	.709	.7749	.6004	.0251	.4915
	STDDEV	2.1904	.146	.7873	.6198	.0194	.6048

II.D.3. Reliability of subsets of items

II.D.3.a. Introduction

A more direct estimate of the effect on test reliability of differences in the number of answer choices on a multiple-choice test* may be obtained from subtests of the various test administrations where items within a subtest are more nearly alike in structure. In addition to structural similarity, the items have been divided into two groups on the basis of whether or not they are problems. As mentioned in the previous section, this is a division according to factor structure, with the subtests of problems possessing a strong unitary factor and the nonnumerical questions (abbreviated 'qual') possessing a complex mixture of weaker factors.

For tests as short as fifteen items, calculated reliability coefficients are considerably lower than the values expected for longer tests.** Coefficients for subtests with as few as five items are very sensitive to the actual distribution or scatter of scores on the two tests. For the extreme case of a subtest of only one item, the distribution of scores is represented by the item difficulty, $\bar{p}$. In this extreme case, the product-moment correlation coefficient between two items (phi, ϕ) cannot reach the value 1.00 unless the item difficulties are identical.

*A short-answer question is considered in this context as a multiple-choice item with an infinite number of response options and a chance score of zero.

**According to the Spearman-Brown prophecy formula, a fifteen-item test with a reliability of 0.65 is equivalent to a sixty-item test with a reliability of 0.88.

As the number of items (k) increases, the distribution of responses approaches some binomial distribution. For tests with many items, the distributions tend to be similar unless the means are quite different. This sensitivity of short subtests to score distributions means that these subtests may seriously underestimate the reliability coefficient even when test means are equal.

II.D.3.b. Results

In order that direct comparison of the various statistics may be made, all statistics have been adjusted to a standard ten-item test length. Two equations were used to adjust for test length. The Spearman-Brown prophecy formula is the widely accepted equation for calculating the new reliability of a test based on a change in its length. The only assumption made when applying this equation is that the items added or deleted are equivalent -- that is, that they have the same intercorrelations, difficulties, and content as the original set of items. The prophecy formula has been used to adjust the correlation coefficient, r. The variances of these subtests were also adjusted for test length by back-calculation from the internal consistency coefficient (KR_{20}). The assumption was made that the coefficient of internal consistency, based as it is on item difficulties and intercorrelations and total test variance, is as appropriate as the Spearman-Brown formula to adjust for test length. Tables of individual subtest statistics are listed in Appendix D.

The reliability coefficients and confidence intervals for subtests are presented in Figures 2.3, 2.4, and 2.5. As can be seen from the overlap of confidence intervals, many coefficients are not significantly different from one another at the $\alpha = .05$ level. Even Codes A1, A2, and A3 -- which are for groups near three hundred -- have intervals spanning more than ten points on the reliability scale.

The important trend in these data is apparent when reliability is plotted against the item chance score (Figure 2.6). The item chance score (ICS) is the reciprocal of the number of answer choices per item and ranges from 0.000 forshort-answer items to 0.250 for four-option multiple choice items. The linear relation between reliability and item chance score is strongest for subtests of problems, weakest for nonproblems, and intermediate for tests of mixed item types. The correlation between reliability and item chance score is 0.82 for problems, 0.33 for nonproblems, and 0.74 for unparceled tests scaled down to the ten-item standard length. The relationship between reliability and item chance score is in the same direction and of approximately the same magnitude for all types of items. However, the strength of the relation is inversely related to the complexity of the factor structure; problems have a strong mathematical factor and hence a more precise correspondence between test reliability and opportunity for guessing.

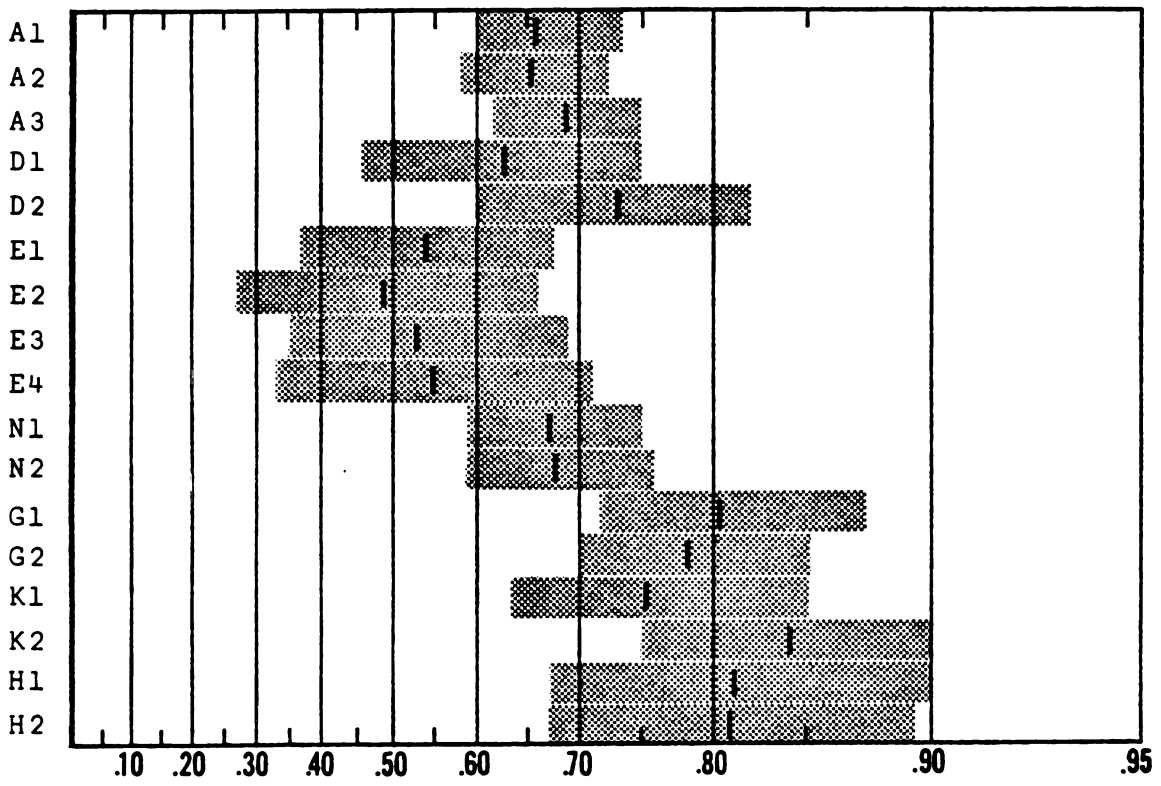


Figure 2.3 Reliabilities and $\pm 2\sigma$ confidence intervals for subtests of problems listed by test code (see Appendix D).

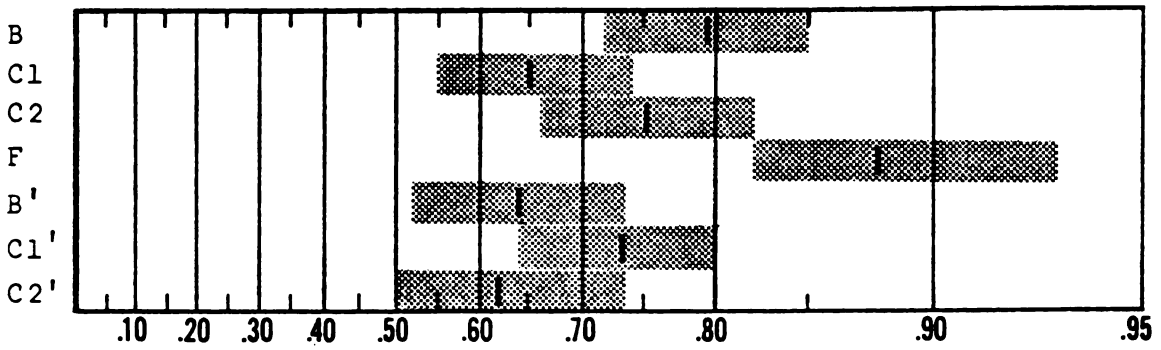


Figure 2.4 Reliabilities and $\pm 2\sigma$ confidence intervals for subtests of identical content listed by test code (see Appendix D).

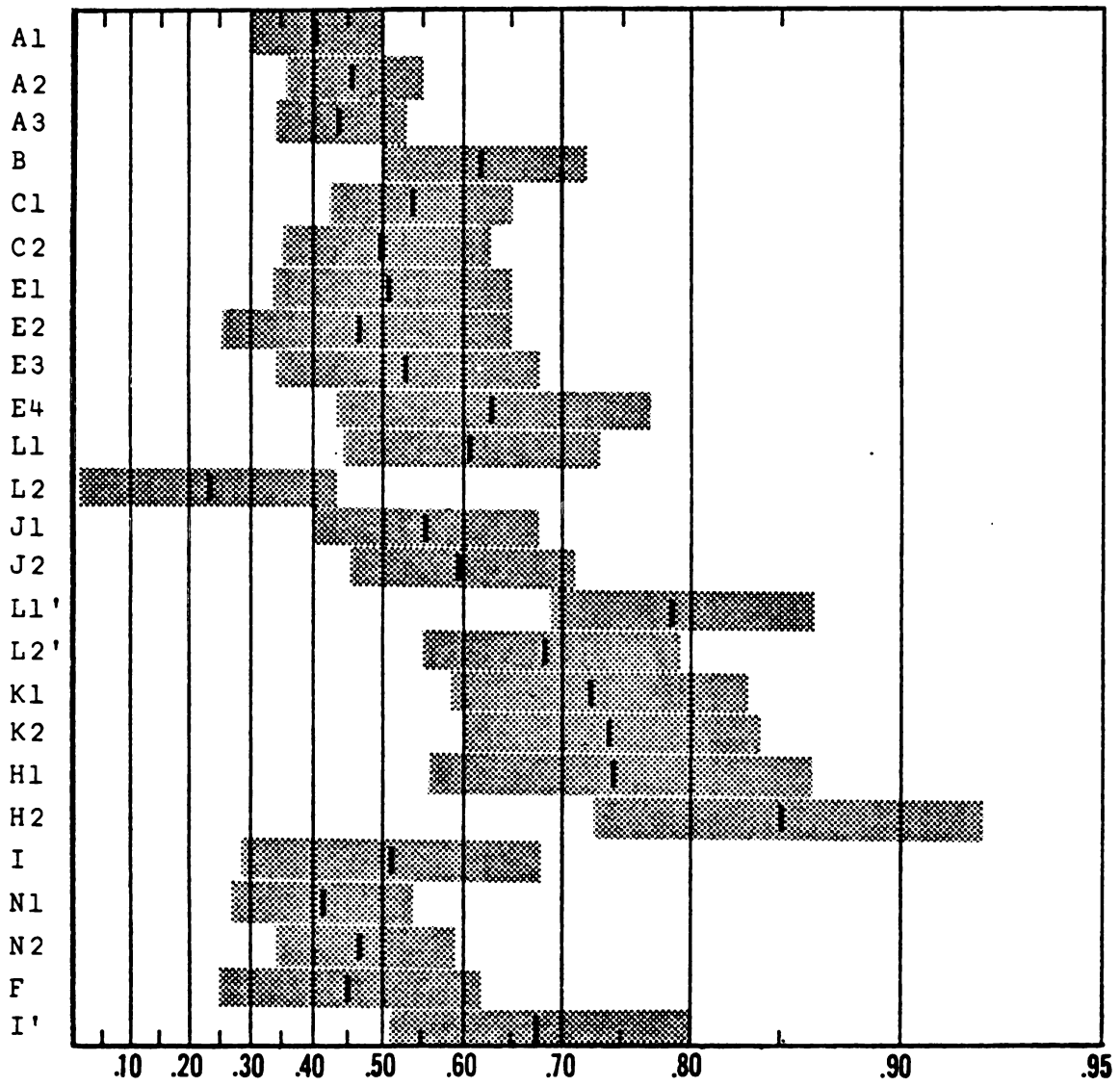


Figure 2.5 Reliabilities and $\pm 2\sigma$ confidence intervals for subtests of nonproblems listed by test code (see Appendix D).

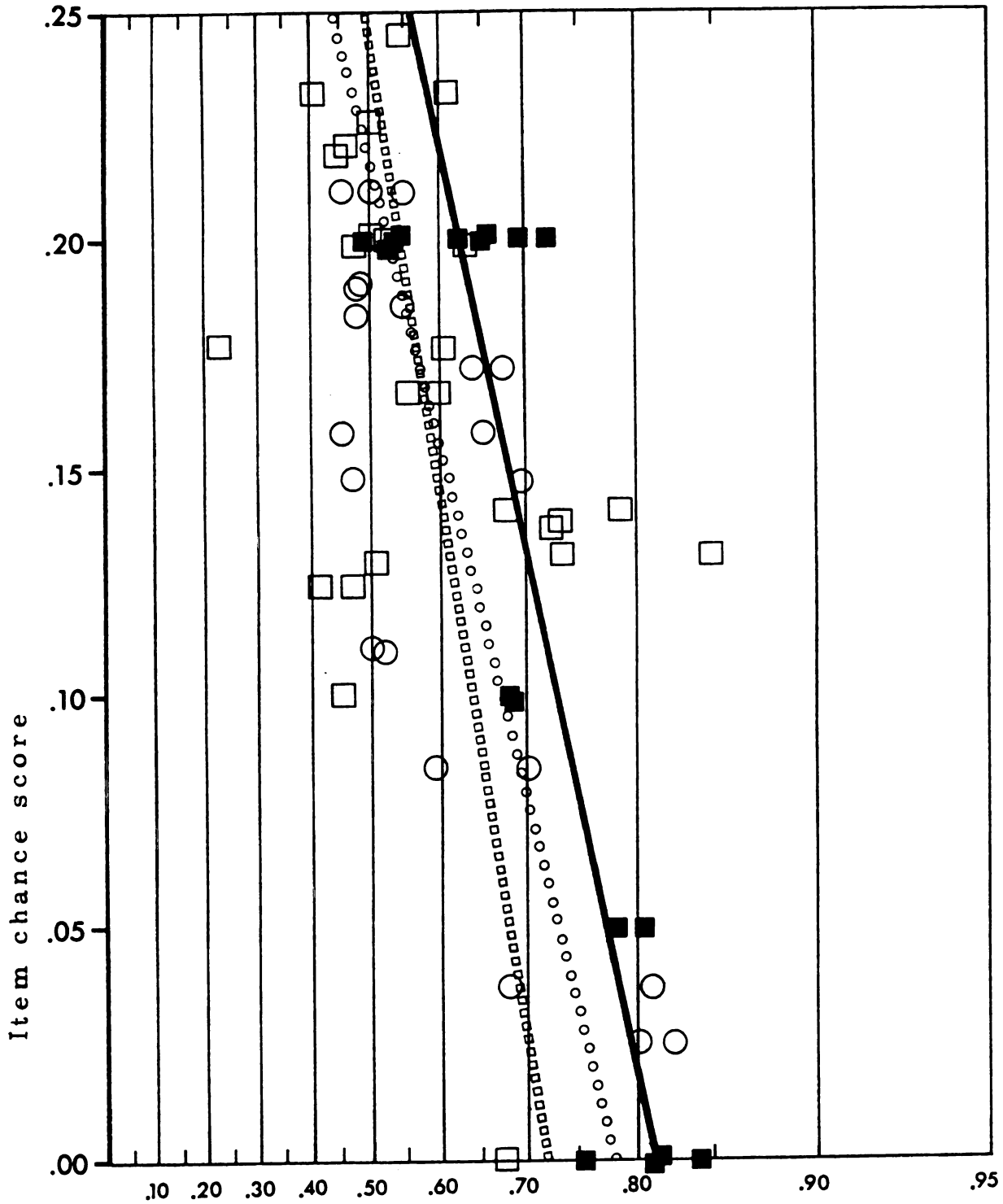


Figure 2.6 Reliability coefficient vs. item chance score for subtests of problems [■], subtests of nonproblems [□], and tests of mixed item types [○]. Corresponding linear equations from Table D.9 are plotted for subtests of problems [▬] and nonproblems [▫▫▫], and fifteen-item tests [ooo].

II.D.3.c. Components of variance

The variance of a sum is not the simple sum of the variances. The variance of the sum of two scores will be larger than the sum of the two variances to the extent that the scores intercorrelate. The equation for calculating this summed variance is

$$s^2_{a+b} = s^2_a + s^2_b + 2 r_{ab} s_a s_b \qquad [2.13]$$

For a sum of n variances there will be $\frac{n(n-1)}{2}$ covariance terms based on the intercorrelations of the variances.

Only when the correlation between variables is zero is the variance of a sum equal to the sum of the variances. In this case, the total test variance may be distributed into several uncorrelated categories whose simple sum is the observed variance. Each of the categories contributes a certain fraction to the total variance. This fraction is the ratio of the component variance to the total observed variance. Thus while both the variances and variance fractions (VF) are additive, the variance fractions possess the additional property of always summing to a mathematically convenient 1.00. The reliability coefficient is simply the variance fraction for true-score variance, and measurement error is whatever fraction and variance remains.

Four categories of error variance were investigated: fluctuation, guessing, equivalence, and memory. Variance ascribed to fluctuation in true scores is significant in meaning but not in size, and variance

due to memory is significant by definition in these pages only for the test-retest situation. Variances attributed to nonequivalence and guessing are both large and important. Each of these four categories will now be treated individually.

II.D.3.c.i. Error variance due to fluctuation

Error variance due to fluctuation in true scores is the result of a change in what the student knows during an exam or between two exams. The magnitude of the fluctuation may be estimated by varying the time between tests and observing the effect of this time variation on the reliability coefficients of the tests. Correlation coefficients for tests with a mean separation time of one hour and a mean separation time of two hours are listed below in Table 2.5.

Table 2.5 Correlation coefficients and the fraction of variance due to fluctuation in true scores (VF_{fl})

Code	r for 1 hour	r for 2 hours	VF_{fl}*
W	.6249	.5871	.0378
X	.6666	.6510	.0156
Y	.5462	.5268	.0184
mean	.6125	.5886	.0239

*$VF_{fl} = r_{1\ hr} - r_{2\ hr}$

Approximately 0.02 unit in reliability is lost to fluctuation in true scores because of the additional hour between test administrations. If the relation is linear, then the reliability coefficients estimated by the correlation between immediately successive tests are also too low by roughly 0.02 unit. This variance fraction due to fluctuation is however only based on five-option multiple-choice tests and may not remain constant for other types of items. Since this source of error is small in any case, it will hereafter be included without identification in residual error.

II. D. 3. c. ii. Error variance due to guessing

The chance success sometimes gained by guessing the answer to a question increases error variance. The fraction of the variance resulting from guessing depends on both the opportunity to guess supplied by the limited number of possible responses and on the motivation to guess when the student is unsure of his answer. The student who knows the answers to all the questions doesn't guess on any of them and the student who gets few questions right probably guessed the answers to many of them. Thus the observed score is the best available estimate of the actual amount of guessing which occurs.

The formula derived by Nunnally [68] to adjust the reliability coefficient for guessing is based on both the estimated amount of guessing (the test mean) and on the chance of random success for a guess. His formula cleanly separated guessing variance from other

sources as illustrated in Figures 2.7, 2.8, and 2.9. As can be seen, the linear equations describing the relation between item chance score and the variance ascribed to guessing are better than the equations based on total error variance. In each case the guessing curve parallels the total error curve, but the points on which it is based are scattered less widely and the coefficient of determination* is much higher. These coefficients are listed below in Table 2.6.

The relationship between guessing opportunity (item chance score) and guessing error is strong and stable for subtests of problems whereas it is much weaker and more variable for subtests of nonproblems. This difference in the strength of the relation between theoretical guessing opportunity and observed guessing error occurs because of the difficulty one has in writing test items which are equivalent in content and difficulty.

Table 2.6 Coefficients of determination for regression equations that predict guessing and error variance from item chance score

Type of data	R^2 guessing	R^2 error
Subtests of nonnumerical questions	0.67	0.11
Subtests of mathematical problems	0.98	0.67
Fifteen-item tests of mixed format	0.95	0.55

*The coefficient of determination is equal to the square of the correlation between the dependent and independent variables in the regression equation. It ranges from zero to unity and is an estimate of the 'goodness of fit' of the data points to the linear equation.

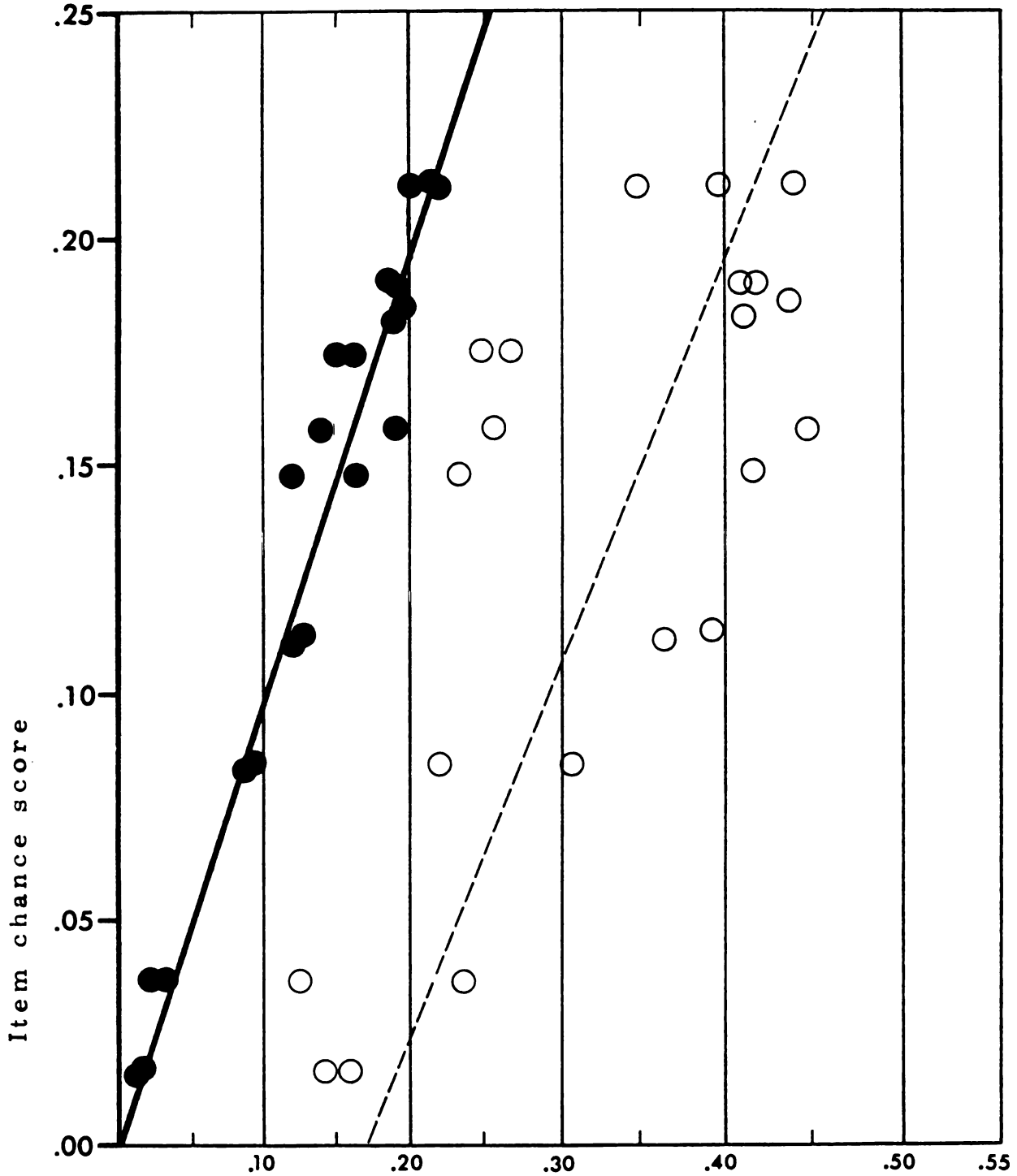


Figure 2.7 Item chance score vs. variance fractions of guessing [●] and total error variance [○] for fifteen-item tests

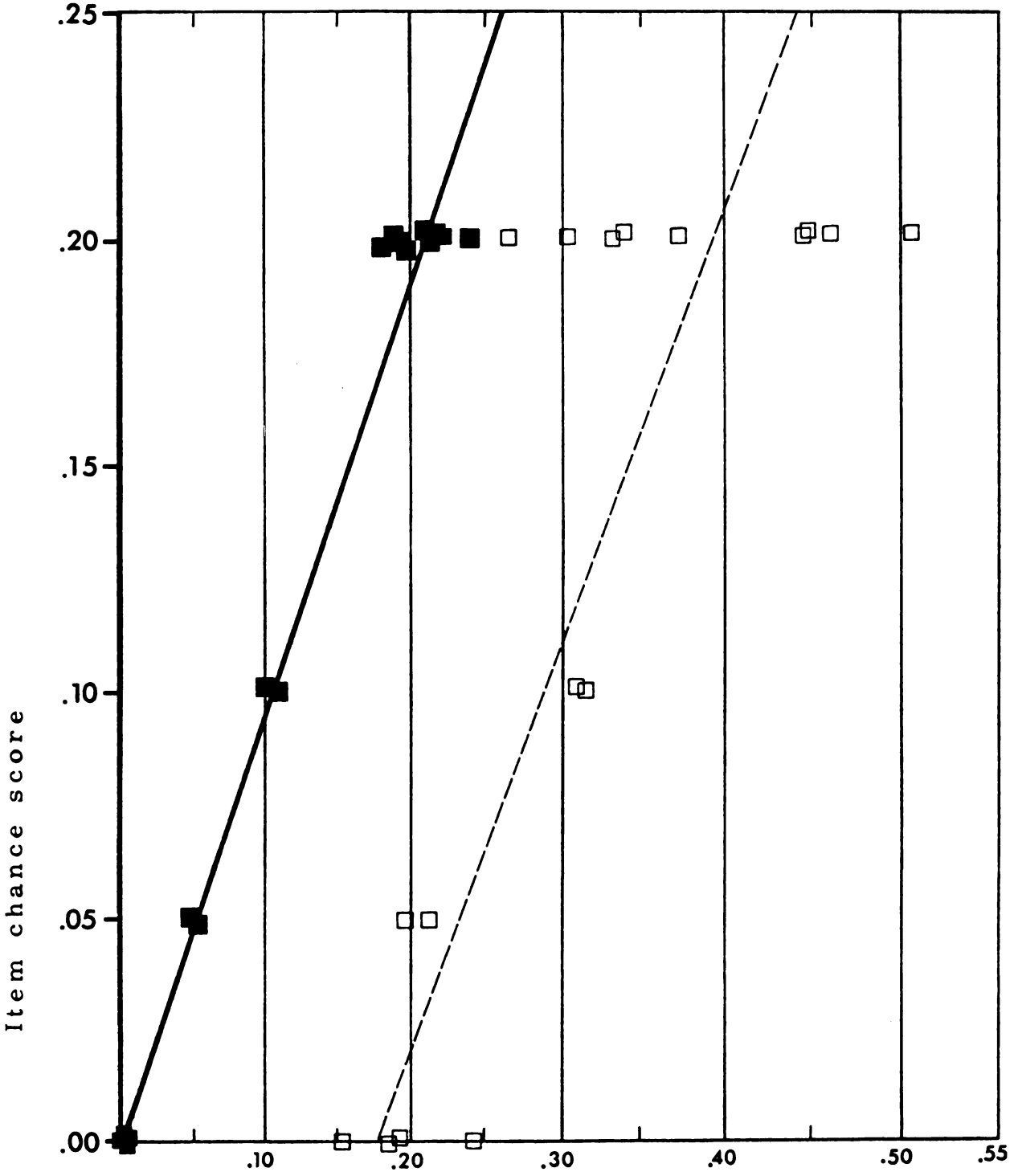


Figure 2.8 Item chance score vs. variance fractions of guessing [■] and total error variance [□] for subtests of problems

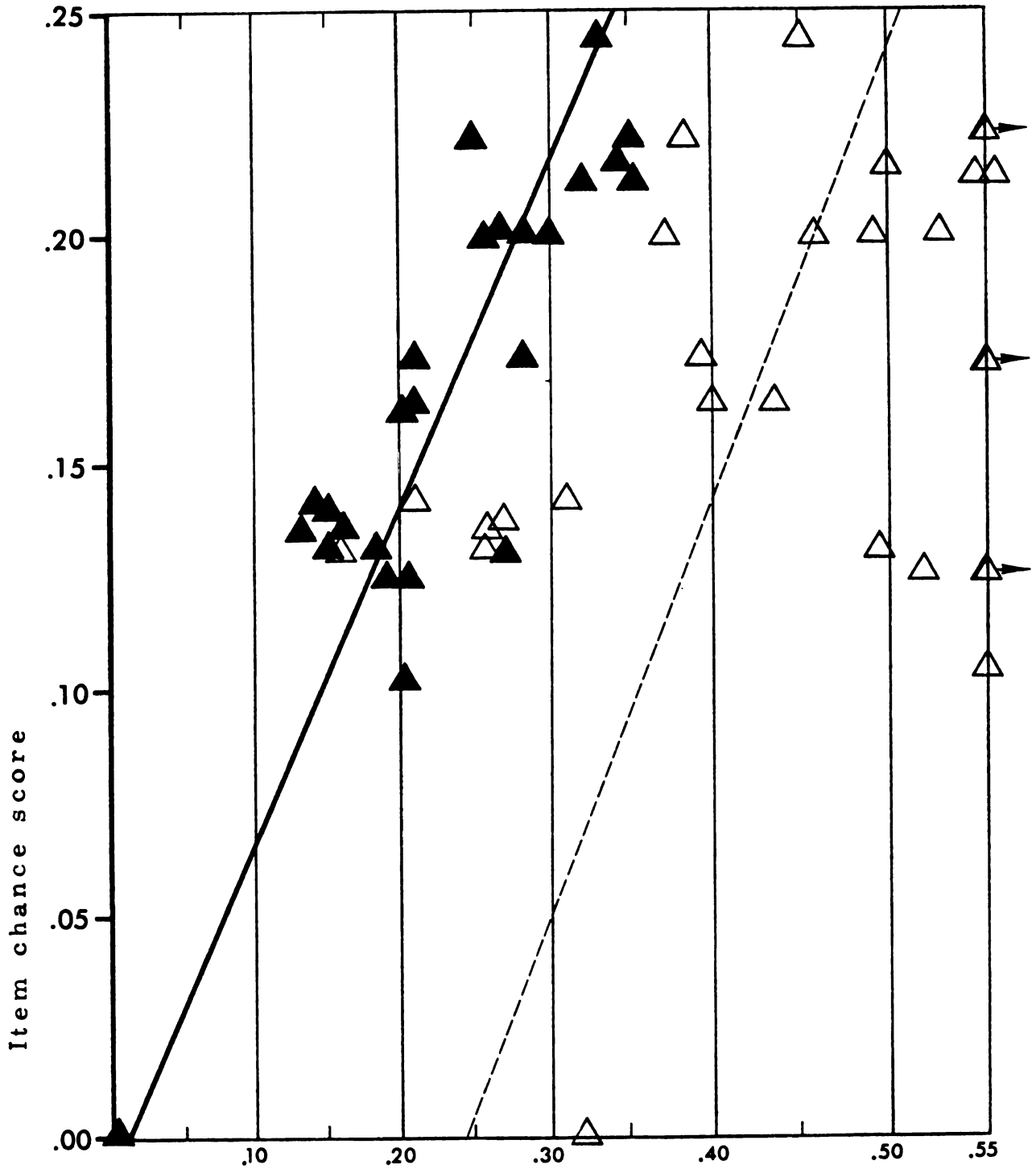


Figure 2.9 Item chance score vs. variance fractions of guessing [▲] and total error variance [△] for subtests of nonproblems

A student who successfully answers one question may fail a supposedly parallel item for essentially two reasons. First, he may truly know the answer to one and not the other. Two questions which test for unconnected bits of knowledge in the same narrow topic very often behave so. An obvious example is the identification of elements or compounds. A student may know the name for $K_2Cr_2O_7$ quite independently from OsO_4. In contrast, parallel problems are seldom completely unconnected since the solution to a problem is based on a mathematical process common to both items. As a result, it is usually easier to write equivalent problems than to write equivalent nonnumerical questions.

The second reason a student may not score the same on two parallel items is that he might have guessed the answer to one or both questions. Clearly, when a question has ten attractive answer options, it is unlikely that the student would get both or even one of the questions right simply by guessing. Yet when a student can eliminate many options from consideration before he chooses, his chances of making a successful guess are much enhanced. It is much more difficult to devise distractors for nonproblems than for problems. It is also more difficult in the case of nonproblems to judge beforehand whether a distractor is ineffective or successful. As a result, the number of effective distractors varies more for nonproblems than for problems.

The differences in variabity of equivalence and guessing errors are illustrated by the coefficients of determination. The differences in

magnitude of equivalence and guessing are illustrated by the variance fractions for these two sources listed in Table 2.7 below. Three additional conclusions may be drawn from these data and other data presented in Appendix D.

Table 2.7 Calculated variance fractions for different types of items

Type	Five-choice guessing	Ten-choice guessing	Short-answer guessing	Residual*
Problems	.2104	.1052	.0000	.1184
Nonproblems	.2748	.1435	.0122	.2446
Mixed types**	.2265	.1157	.0048	.2192

*Residual error is almost entirely that of nonequivalence between forms
**These values are for tests of fifteen items adjusted to the standard ten-item test length of the other data in the table.

First, the relative contribution of guessing error to total error variance is directly proportional to the chance of making a successful gues. The relation is weakest -- though still strong -- for nonproblems because of the unpredictable effectiveness of nonnumerical distractors. Second, the magnitude of the true-score variance is inversely related to guessing opportunity. In contrast with the effect of guessing on error, the effect of guessing on true-score variance is not linear. True-score variance increases as the opportunity to guess decreases, but the magnitude of the increase exceeds the concomitant decrease in error variance. Third, the total error variance does not decrease by the same amount as does the guessing error decrease. The residual

error of nonequivalence tends to increase slightly as the guessing error decreases so that the observed decrease in total error is less than expected. Total test variance is the sum of true-score and error variance, and since true-score variance increases more than expected while error variance decreases less than expected when guessing opportunity is reduced, the overall effect is to significantly increase total test variance.

In sum, as the guessing opportunity decreases (1) guessing error variance decreases in direct proportion, (2) true-score variance increases faster than expected, and (3) total error variance decreases more slowly than expected. These nonlinear effects even out so that the reliability is linearly related to the item chance score.

II.D.3.c.iii. Error variance due to nonequivalence

The method of estimating the reliability of tests by parallel forms correlation introduces a major possible source of error variance.* Two different but equivalent forms are used to estimate the average precision of each by intercorrelation, and any extent to which the two forms differ in content will lower their intercorrelation. This forms effect discussed in Section II.D.1.c.ii. is caused by the lack of equivalence between the parallel forms of an examination. After

*This error is not an artifact of the method of estimating reliability unless one is interested in the specific reliability of a particular test form and not in the reliability of a class of tests from which the ones being used are sampled.

the separation of guessing variance from total error variance, the bulk of the remaining error is attributable to this forms effect. As can be seen in the graphs of item chance score versus error variance (Figures 2.7, 2.8, and 2.9), the variance fraction of nonequivalence is approximately constant across the scale. However, this residual fraction of error variance does change with the content or factor structure of the test.

Subtests of problems are most equivalent (least amount of nonequivalence), nonnumerical questions least equivalent, and tests of mixed types are of intermediate equivalence. Nonequivalence quickly surpasses guessing by default as the single greatest source of error variance when the number of answer choices is increased. When tests with ten choices are used, the error due to guessing is about half that due to nonequivalence. Total elimination of nonequivalence between parallel forms of a test would have almost as great an effect on reliability as total elimination of guessing error. Although this is theoretically possible in some instances, in practice the effort needed to eliminate the last vestiges of nonequivalence would far outweigh any marginal improvement in reliability. Test reliabilities of 1.00 are the ideal by which real tests are judged; only rarely will perfection be closely approached.

II.D.3.c.iv. Error variance due to memory

This last category of error to be investigated is of a special nature limited to few actual testing situations. As it is defined, error variance due to fluctuations in or because of memory can only exist when the same examination is given twice to the same students. However, when the same questions can be selected from a computer bank to appear on more than one examination, there is a good chance that students will now and then encounter questions they have met on previous tests. The total variance of any test is the complex sum of the variances of the individual items, and if a particular item is met again, memory variance will contribute to error variance for that item, albeit only a small fraction of the total variance.

A rough estimate of the importance of memory variance may be gleaned from the study of error and guessing variance fractions of short-answer versus multiple-choice questions. These variance fractions are plotted against the harmonic mean item chance score in Figure 2.10. The curve for guessing is similar to guessing curves for other data, but in this case, the residual error is predominantly memory variance since both tests in each comparison have identical content. It may be that there is a measure of nonequivalence introduced by changing the format of an item from short-answer to multiple-choice, but the apparent identical nature of the items used in these comparison leads this author to conclude that it is not substantial. If nonequivalence were introduced, then the curve extrapolated to zero

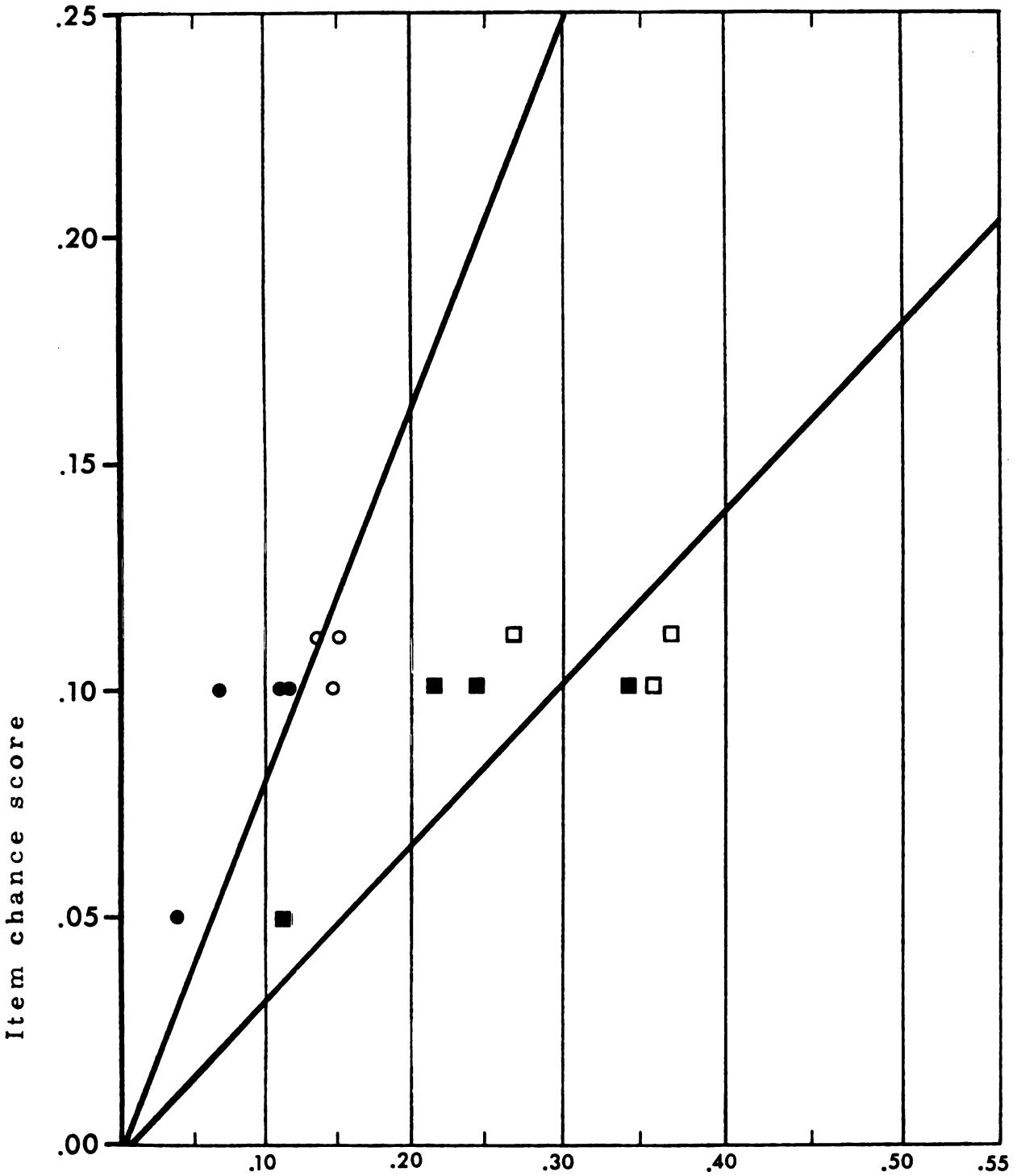


Figure 2.10 Item chance score vs. guessing [●,○] and error [■,□] variance for subtests of short-answer items compared with identical multiple-choice items.

would have a sizeable positive intercept. The intercept is 0.0285 units and the deviation from zero is al likely to be caused by the small number (seven) of points on which the equation is based as on any induced nonequivalence due to the change in item format.

Table 2.8 Calculated variance fractions due to guessing and memory

Type	Five-choices	Ten-choices
$VF_{guessing}$	.34	.19
VF_{memory}	.30	.15

The variance fractions for guessing and memory are listed in Table 2.8. Memory is about four-fifths as important as guessing, and decreases in the same manner as the number of response options is increased. Still, errors of memory variance will be trivial if only one item on an exam is repeated. Not more than two percent of the total variance and probably much less will arise from memory.

II.D.4. Type of item structure and variability in test means

The product-moment correlation coefficient is a measure of the reproducibility of a distribution relative to its mean. When the distribution is effectively unbounded and no scores 'pile up' at either extreme of the scale, then the correlation between two distributions can be unity even when the means are different. A new statistic was

devised by the author to describe this instability in test means, since this instability is not quantified by any other available statistic.

The best estimate of the true mean based on two equivalent tests given to the same group of students is simply the average of the observed means. In this manner, each of a pair of tests deviates from its 'true' mean by half the difference between the observed means. These deviations from the estimated true mean may be used to calculate a stability estimate of the mean according to

$$\text{SEM} = \sqrt{\frac{\sum\left(\frac{d}{2}\right)^2}{n - 1}} \qquad [2.14]$$

where d is the observed difference in test means between two tests given to the same group of students and n is the number of paired tests in the set of data.

The stability estimate of the mean is interpreted as is any standard error of measurement; the larger the SEM the more unstable or imprecise the measurement. The stability estimates for three groups of test means are listed in Table 2.9 on the next page. The tests are grouped according to the dominant item-type: four- and five-option multiple-choice items, six- to ten-option items, and short-answer items.

The relation between the item chance score and the stability estimate of the mean is weak and curvilinear. The linear regression equation is $^{d}/_{2} = 3.0743 - 5.9045\,(\text{ICS})$ and the correlation is only 0.24.

Table 2.9 Stability estimate of the mean (SEM) for groups of tests*

Groups	ICS**	KR_{20}	n of pairs	$^{d}/_{2}$	SEM
A, D, E	.193	.6887	9	2.26	3.06
L, J, N	.139	.6405	6	1.73	2.38
H, K	.032	.7710	4	2.93	3.58

*In this table, $^{d}/_{2}$ and SEM are expressed in units of percent.
**Item chance score, the reciprocal of the harmonic mean number of answer choices per question

The individual differences between pairs of tests are tabulated in Appendix C, Table C.1. There is about as much variation in SEM within groups of tests as between groups. However, if these results are not peculiar to this small collection of data, two conclusions may be drawn. First, there is less variation in difficulty from test to test as the number of answer choices increases. Second, short-answer items depart from this trend and show the greatest variation in difficulty from test to test. The conclusions may be explained as follows.

The mean difficulty of multiple-choice items with few response options is subject to much fluctuation because a small number of poor distractors has a sizeable effect on the relative guessing success and hence the item difficulty. For example, compare a five-option item with a ten-option item. When all distractors are at least somewhat effective, the item chance scores are 0.20 and 0.10 respectively. When three distractors are ineffective for each item, the respective

item chance scores are 0.50 and 0.14. The five-option item is more than twice as susceptible to guessing whereas the ten-option item is only about forty percent more susceptible.

The short-answer item does not continue the trend. Although the item chance score is zero for the short-answer item, there can still be considerable variation in intrinsic difficulty between supposedly parallel items. It is much easier to judge the apparent difficulty of an item in the multiple-choice format. The universe of possible answers to a question and the required fineness of distinction between right and wrong responses are controlled by the answer choices provided by a multiple-choice item, while these are often uncontrolled or undefined for a short-answer item. What appear to be equally difficult items may in practice be items quite different in difficulty.

II.D.5. Standard error of measurement for tests and subtests

The standard error of measurement is calculated from the reliability coefficient and the standard deviation according to

$$SEM = s_t \sqrt{1 - r} \qquad [2.3]$$

Lord [34] showed that this value is the mean standard error based on sampling theory. The 'reliability coefficient' in this case is the KR_{21} coefficient of internal consistency which is a lower bound to consistency and hence gives an upper bound estimate of the standard error. However, the internal consistency is not a good estimator of

reliability for short stratified tests such as those used in CEM 130/131. In this case a more valid standard error is calculated from the parallel forms correlation reliability coefficient. The values for the standard error of measurement are plotted against item chance score in Figure 2.11 on page 140.

The relationship between guessing opportunity and probable error in test scores is weak and unstable.* If guessing were the major source of measurement error there would be a dramatic decrease in the standard error when item chance scores approach zero as for short-answer items. However, the mean decrease in standard error as the number of response options increases from five to infinity is often less than the difference in standard errors between tests with similar numbers of response options. There are also several instances where tests composed of mainly short-answer questions have larger standard errors of measurement than tests composed of five-option multiple-choice items. Thus measurement error is not strongly related to guessing error, whereas in contrast reliability does depend significantly on guessing error.

The maximum difference between any value of the standard error and the mean value from all tests is about the same as the differences

*The regression equations and coefficients of determination describing this relation are listed below. Because the coefficient for subtests of nonnumerical questions was so low, these points are not plotted.

Fifteen-item tests	SEM = 1.3319 + 1.9444(ICS)	$R^2 = 0.30$
Subtests of problems	SEM = 1.0782 + 0.9466(ICS)	$R^2 = 0.31$
Subtests of nonproblems	SEM = 0.9592 + 1.0416(ICS)	$R^2 = 0.06$

predicted from regression or the differences within a single group of tests with similar item chance scores. This variability in standard error is tabulated in Table 2.10.

Table 2.10 Variation in standard error of measurement between and within guessing levels*

Variation	Fifteen-item tests	Subtests of problems
Within levels	21 to 28 percent	16 to 31 percent
Between levels	27 percent max.	18 percent max
From grand mean	22 to 28 percent	18 to 25 percent

*A guessing level is a group of tests with similar item chance scores

The standard error could be predicted about as well by simply taking the overall mean as it could from a regression equation. Apart from what may be considered as random fluctuations (or fluctuations related to some unspecified characteristic) the standard error is approximately constant. This conclusion was also empirically demonstrated by Lord [34]. He found that the square root of the number of test questions adequately described the standard error of measurement. Given the proportionality SEM $\propto \sqrt{n}$, it is easy to calculate that to decrease the standard error for fifteen-item tests by half,* the tests would have to be increased to sixty items.

*A typical standard error for a fifteen-item test is 1.5 score units or ten percentage units. If the test is increased to sixty items, the standard error increases to 3.0 score units but decreases to only five percentage units. Thus the relative size of the error is cut in half.

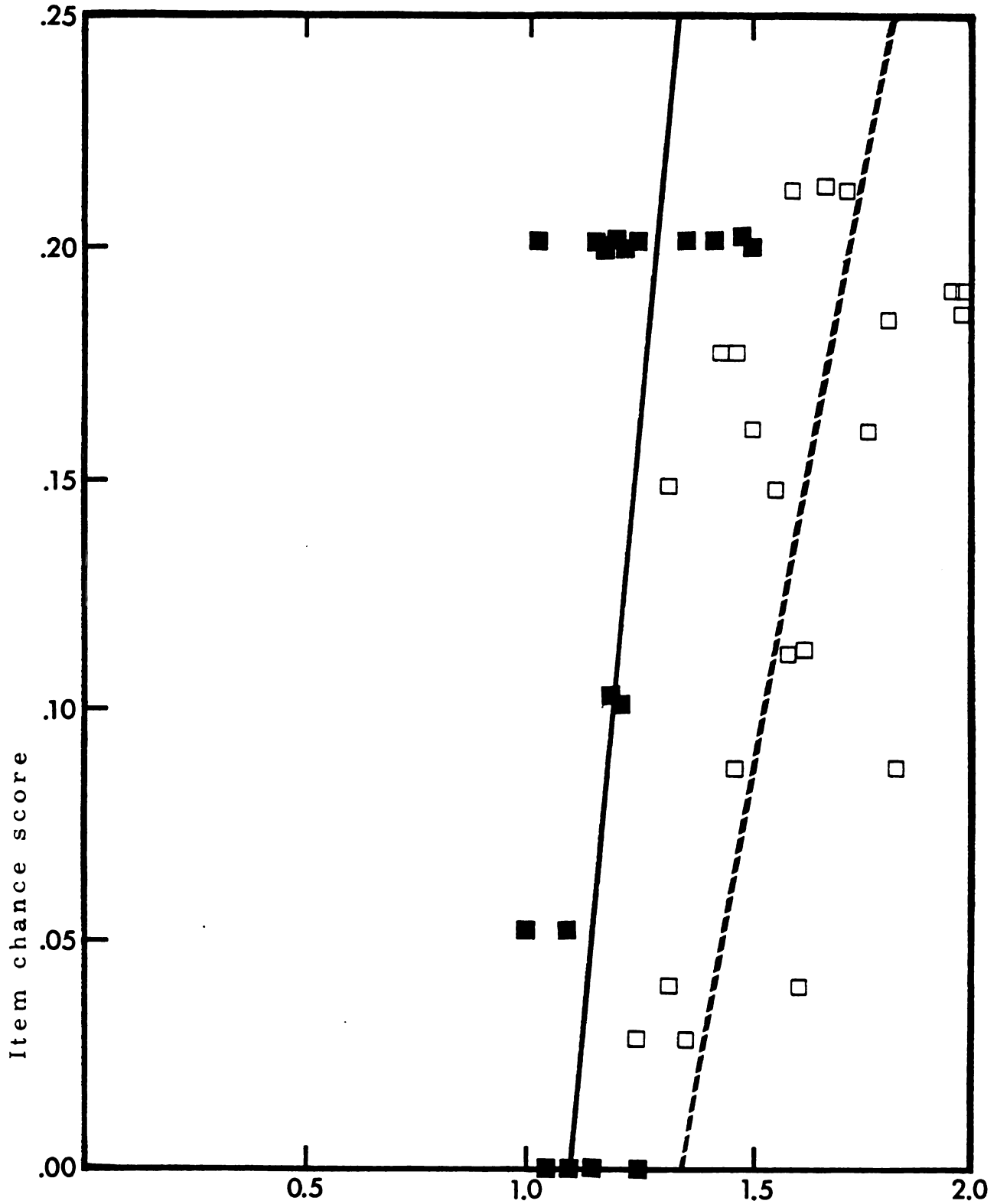


Figure 2.11 Standard error of measurement vs. item chance score for fifteen-item tests [□] and ten-item subtests of problems [■]

II. E. Summary and conclusions for errors in measurement

The major conclusions of this research in measurement error are summarized below in fourteen short statements. Following this list, the practical implications of these conclusions will be discussed. Lastly, a series of general observations and recommendations to the test constructor will be made.

II. E. 1. List of research conclusions

1. Guessing error is directly proportional to the fractional probability of guessing correctly. Guessing error is halved by increasing the number of answer choices from five to ten.

2. The number of answer choices was found to be the most important of eight variables in the prediction of test reliability from linear regression. It accounts for thirty-four percent of the variation in reliability coefficients. The residual or unexplained variance was thirty-eight percent and includes differences in equivalence between pairs of correlated tests.

3. Parallel forms of an exam are not statistically identical. This lack of equivalence in content between different tests of the same topics introduced error often greater than that caused by guessing.

4. Residual error including nonequivalence is the same size as guessing error for six-choice items. It remains approximately constant and does not decrease as guessing opportunity decreases.

5. Examinations consisting of problems tend to be more reliable than those consisting of nonproblems. This difference in reliability is small (seven percent) for short-answer questions and substantial (twenty-five percent) for five-choice questions.

6. The reliability of tests consisting of nonnumerical questions is more unpredictable than that of tests consisting of problems. The spread in reliability coefficients for nonproblems is double the spread for problems. It is possible for a test of nonproblems to have an unusually high reliability quite by accident.

7. The 'probable error' of measurement tends to decrease slightly as the guessing opportunity decreases, but deviations in probable error from other sources are as large or larger.

8. A short-answer test item may be carefully rewritten as a multiple-choice item without changing what the item really tests, although the difficulty of the item may change considerably.

9. When the same test questions are encountered by the student on an immediate retest, memory variance is four-fifths as large as guessing variance. Memory variance decreases erratically as time between testings increases.

10. It appears that as the number of answer choices increases, tests become more similar in difficulty. Short-answer tests, however, seem to have the least stable means.

11. In these studies, the method of assigning students to take different forms of an examination did not lead to differences between groups. Statistical equality of the groups was achieved by nonsystematically distributing tests to students as they entered the exam room.

12. In these studies, the order in which tests were taken had no effect on the calculated reliability coefficients.

13. Fluctuation in true scores is of minor importance when one or two hours separates test administrations. Reliability decreased only 0.02 units when the time between tests increased from one hour to two hours.

14. When test questions have different numbers of answer choices, the harmonic mean of the number of choices per item is the appropriate method of calculating the 'average' number of answer choices. The short-answer item is treated as a multiple-choice item with an infinite number of response options.

II.E.2. Implications of some important conclusions

Grades in CEM 130/131 are assigned according to a preset scale, yet students take different exams at different times and are allowed to retake a different form of the same exam when they are dissatisfied with their score. Under these circumstances, the precision and accuracy of different forms of the same exam relative to one another

are of special concern. Since the student may be graded on his performance on any one of several parallel forms of an examination, it is important that all the different forms of each exam be truly equivalent in both content and difficulty. Also, since the student may retake an exam several times, it is important that his final score reflect his knowledge and not his guessing luck. The theoretical and experimental relationships between guessing error, test equivalence -- both of content and difficulty -- and the types of items which appear on the test will now be summarized.

Guessing error is probably the single greatest impediment to acceptance of the objective or multiple-choice item. That a student can receive credit for a question without knowledge of the answer is considered a fatally damning indictment of the type of item which allows guessing to occur. The results of this research clearly demonstrate that by increasing the number of answer choices, the effect of guessing can be reduced to the point where the injury done the test reliability may be minor rather than critical. The guessing error expected from four- and five-choice items is more than halved by increasing the number of answer choices to ten.

It is also true that short-answer or fill-in questions are not immune to guessing. In theoretical discussions and calculations it is assumed that the short-answer item may be treated as a multiple-choice item with an infinite number of answer choices. The conclusion that the short-answer item therefore has zero guessing error is valid

only if the possibilities through which the student sifts is effectively infinite. Many short-answer questions do not have an effectively infinite set of possible answers. For example, "How many two-fold axes are there in the molecule... ?" Though it might seem that there are a very large number of possible answers to the completely uninitiated, the student with any exposure at all to this topic quickly realizes that the answer is an integer between zero and twelve. Other short-answer questions have fewer choices and others have more, but very often the effective number of possible answers is low enough to make guessing a non-zero source of error. Thus the recall item too can suffer from guessing as can the recognition item.

Guessing is not constant for all students. The good student guesses less often and with more success than the poor student. This means that there is less error in the highest scores than in the lower scores, and that there is less guessing error on an easy test than on a hard test. The high scores on any test are the most reliable. There is much more chance of giving a student a 1.0 when he should have gotten a 0.0 than there is of giving a student a 4.0 when he should have received a 3.5.

The prediction of actual guessing is based on the expected chance of success and depends on more than just the number of answer choices. The type of question influences the predictability of guessing errors. For example, a division of test items into problems and nonproblems produces a striking difference in the ability to predict guessing error.

It is much more difficult to construct effective distractors for nonproblems, and because of this the reliabilities of tests of nonproblems are more variable. Increasing the number of choices does increase test reliability and decrease guessing error, but the predicted values do not equal those for problems even at infinity.

A second important source of 'error' is the error due to nonequivalence between different forms of the same examination. Nonequivalence is an error in the sense of inaccurate estimation of how much the student knows of what was tested. However, if what was tested is not representative of what could have been tested, then the test score is not representative ofthe student's total knowledge of the course material. A test composed of items from two or three of five topics will misrepresent the student's knowledge unless it is known that the chosen topics are similar enough in content to the untested topics so that the student who answers 'five from column A' could just as readily answered 'five from column E.' Nonequivalence error is analogous to faulty sample selection for analytical analysis. One would not select all the samples from the same region of the specimen unless one knew that the specimen were homogeneous.

The size of the error caused by lack of equivalence is roughly equal to the guessing error of six-choice questions. It does not consistently decrease or increase as the number of response options changes. Accordingly, lowering guessing opportunity, increasing item discrimination, or testing for higher mental processes may

improve the quality of the questions but will not much affect their representativeness or equivalence.

When the number of testable concepts in the topics to be covered on an exam is larger than the number of questions, it is not possible for a single test to cover all concepts. There are three ways of selecting test items for multi-form tests. One way is to prepare a balanced selection of question ideas and construct several parallel items for each idea. Tests constructed in this manner are item-parallel. A second method is to write a number of questions for each testable concept (one for each concept if the number of concepts is large) and randomly select the individual test items from this pool. Tests of this sort are random-parallel. The third method also entails writing a collection of items for all testable concepts. The question ideas are grouped into content categories and test items are selected randomly from among these categories. This last way yields tests which are stratified-random-parallel. The tests used in CEM 130/131 are generated according to this method.

There is less chance for large nonequivalence error when the pool of test questions is stratified before selection than when selection is purely random from the entire pool. Still, there is somewhat more nonequivalence present than if all forms of the test were item-parallel. The item-parallel method is most often used when only two forms of an exam are needed for simultaneous administration to a large group of students. When many forms are needed over a period of days,

students may quickly decipher which concepts do not appear on the tests, so that some random representativeness in the selection of test ideas is necessary.

Regardless of how exams are constructed, their difficulty concerns the instructor who uses them. When different students take different tests which cover the same material, the tests should be of equivalent difficulty. This second kind of equivalence is not related to test reliability unless the distribution of test scores is markedly distorted by a ceiling or floor effect. But whenever students will be graded on different tests according to the same grading scale, the tests should be equally difficult. It appears that short-answer test questions are least stable in difficulty, four- and five-choice items of intermediate stability, and eight- and ten-choice items of greatest stability in difficulty. It also appears that short-answer and multiple-choice items can, with careful rewriting, be interchanged without changing what the items test. Thus equivalence of content can be preserved, and equivalence of difficulty enhanced, if short-answer items are rewritten as multiple-choice items. The guessing error can be held to an eminently acceptable level by employing ten or more answer choices per question. So too can memory errors be held to an acceptably low level when only one or two identical questions appear on a second form.

The traditional yardstick of measurement precision is the standard error of measurement or confidence interval. However, the

standard error is not a good criterion by which to improve test quality. It is best used to evaluate tests which have been improved according to other criteria such as guessing error reduction, increased parallelism, content representativeness, difficulty level, and item discrimination. These other criteria are estimated en masse by the test reliability.

As an illustration of why the probable error is not a satisfactory improvement criterion, guessing opportunity may be compared with both the standard error of measurement and with the reliability coefficient. For tests of problems, guessing opportunity is strongly related to reliability; the expected chance score accounts for 67 percent of the change in reliability as the number of answer choices increases. In contrast, the expected chance score accounts for only 31 percent of the change in standard error. Also, there is only about half as much variation in the probable error as there is in the reliability; the probable error of a test is more nearly constant. Roughly speaking, it's a case of not being able to predict the variation in something which doesn't change much anyway. The best theoretical prediction of the probable error in a test score is based on the raw score and the test length and not on test-quality criteria such as reliability or disctimination.

II. E. 3. List of aphorisms and recommendations

1. The accuracy and precision of a student's grade increases when the reliability of the term average increases. The reliability of this overall assessment of student achievement waxes as the independent measurements -- the tests -- become more numerous or more individually reliable.
   - 1.1 The reliability of a test increases with its length. This is the simplest method of increasing measurement precision. The time available for the administration of the test is the 'limiting reagent' in most cases and the maximum feasible number of test questions for a given class period does not often yield a sufficiently reliable test score.
   - 1.2 The reliability of a test increases as the individual test questions become more reliable. Some practical ways to increase item reliability and reduce random error are mentioned in later recommendations.
   - 1.3 The reliability of the term average increases as the number of tests on which it is based increases. By increasing the number of tests, the single-test restrictions on time and length are circumvented. When the number of examinations increases, each exam covers less material but covers it more reliably and in greater depth.
   - 1.4 Diminishing the scope of a course increases the precision of the tests given in the course. The range of topics to be taught in a course is seldom discretionary; thus this method is impracal. However, those in a position to determine curricular content and course syllabi should note that to require a large mass of material to be presented, mastered, and tested almost ensures that it will be presented hurriedly, mastered tenuously, and tested unreliably.

2. Increasing the number of answer choices reduces guessing error and enhances item reliability.
   - 2.1 Test questions with five or fewer answer choices are too unreliable for short tests. The test score from fifteen five-choice questions consists of about 23 percent guessing error, 22 percent additional error, and only 55 percent reliable measurement. When the number of response options decreases to two -- as for the true-false question -- guessing error balloons to 51 percent and the reliability shrivels to only 27 percent.
   - 2.2 The guessing error of five-choice questions is halved by increasing the number of answer choices to ten. Further doubling the number of choices continues to halve guessing error, but more than twenty answer choices per item is not practicable.

2.3 Fill-in or short-answer items are not immune to guessing. Often the small number of actual choices from which the answer must be selected is readily obvious to even the poorer student. Whenever the student has narrowed his options to some reasonably small number, he can 'guess' the answer even though no explicit prompts are eliminated.

3. Increasing the number of answer choices sometimes increases the time needed by the student to take the test.
   3.1 When the student must carefully weigh each answer against the others, the requisite time to answer the question is proportional to the number of choices. Questions which ask for a 'best answer' or 'best example' are frequently of this type. Eight answer choices is a reasonable maximum for such response-content items.
   3.2 When the student can answer the question in his own mind without refering to the set of answer choices, the time required to answer a question is little affected by the number of choices. Problems with numerical answers are good examples of this stem-content type of item. Items with up to twenty choices may be used profitably to decrease the guessing error without increasing the time of taking a test.
   3.3 The time required to complete a fill-in or short-answer question cannot be estimated beforehand. If there is any doubt in the student's mind about exactly what the question is asking, the time he spends on the question increases. The set of answers provided by the multiple-choice question helps the student grasp the content of the item. Occasionally the observed difficulty of a short-answer question may be due to its ambiguity.

4. Multiple-choice problems are more reliable than nonproblems. The greater precision of a test composed of problems results from the conjunction of a strong factor structure and more uniform item quality.
   4.1 The many different examples of chemical problems all have in common their underlying mathematical nature. Algebra and arithmetic are superimposed on the chemistry of the questions and the effect is to narrow the breadth of the topics. Nonproblems lack this reinforcing relatedness.
   4.2 Distractors are easier to write for problems than for nonproblems. When ten or more common-mistake wrong answers are not available, it is a simple matter to round out the number of answer choices with selected 'random' numbers. In contrast, it is often very difficult to construct plausible distractors for nonproblems. Thus the guessing rate is more easily lowered for problems than nonproblems.

4.3 Distractors for problems are not easily dismissed as obviously wrong. However, some distractors for nonproblems may be promptly dismissed as obviously wrong by most students. Thus the actual guessing success on a nonproblem may be much higher than the theoretical success rate, and it may vary considerably from student to student. The chance success rate for problems is in practice much more uniform and predictable for students and for questions.

5. The difficulty level of a multiple-choice exam is somewhat more stable and predictable than that of a short-answer exam. Because the level of discrimination needed to distinguish the correct from the incorrect can be more easily controlled in the multiple-choice format, the difficulty level is more easily predetermined.

6. Most short-answer questions can be rewritten as multiple-choice questions without changing what the items test.
   6.1 Unless many response options are used, the multiple-choice item will be significantly easier than the short-answer item.
   6.2 A short-answer item which results in a seemingly trivial multiple-choice item is probably already a trivial question in the short-answer format.
   6.3 It is more often the case that a multiple-choice item cannot be rewritten in short-answer format than the reverse. A frequent example of such intrinsically multiple-choice questions asks the student to select the answer which is most basic, least ideal, highest boiling, and the like. This kind of question must have a limited, well-defined set of possible answers as provided by the answer key. Only when the defined set of responses has an unambiguous class name such as colligative property or halide can such a question be asked in short-answer format.

7. All test questions should be similar in difficulty. The only time test questions must cover a range of difficulties is when all the questions are from the same narrow subtopic. For the typical range of topics on a chemistry test, the best overall discrimination between students across all grading levels is achieved by writing each of the test questions to be of medium difficulty.
   7.1 The range of topics on an exam is estimated by a coefficient of internal consistency. The appropriate coefficient for this purpose is the Kuder-Richardson Formula 21 (KR_{21}). When the KR_{21} is not near 1.00, then the difficulty levels of all test items should lie between 0.40 and 0.70.

7.2 When teaching and testing are for complete mastery, and grades are based on a pass-fail criterion, the consideration of item difficulties is not relevant. What matters is whether getting a question right indicates complete mastery of the concept.

8. The probable error of a test score is smallest for high scores regardless of the test mean. Consequently, the precision of a test can be concentrated in different areas of the score distribution by controlling the test difficulty.
   8.1 Difficult tests differentiate most precisely among the best students. If the instructor desires only to define the highest rank of students, a very difficult test should be written. Many students will score zero, but the instructor does not in this instance wish to distinguish different degrees of failure. It is particularly important on a difficult test that item discrimination be uniformly high and guessing error very low. Recall items are best for the difficult test.
   8.2 Easy tests differentiate most precisely among the poorest students. If a minimum acceptable level of achievement is the criterion by which students will be judged, an easy test should be written. Many students will score perfect, but the instructor desires in this instance only to discriminate between adequate and inadequate achievement.
   8.3 Medium-difficulty tests provide the best average discrimination across all levels of achievement. Precision at the bottom is not wholly sacrificed to improve it at the top, as on the extremely difficult test, nor the converse as on the very easy test.
   8.4 Increasing the length of the test increases precision at all levels of achievement regardless of score distribution.

9. The effort required to attain a balanced and representative test is a function of the factor structure of the topics and concepts presented in the course. The KR_{21} is an indirect measure of the content homogeneity of an examination.
   9.1 When the material to be covered on a test is homogeneous, simple random selection of specific questions for the test provides a representative sample of student achievement.
   9.2 When the material to be covered can be clustered into homogeneous segments, stratified-random sampling provides the most representative sample of test questions.
   9.3 When the material to be covered is heterogeneous and does not contain homogeneous subtopic divisions, no procedure can assure that a small sample of items will adequately represent the material in the course. This is unlikely to be the case for the typical chemistry course.

10. When students do not all take the same form of an examination, all forms of the exam should be equivalent in content and difficulty. The procedures used to achieve content equivalence are the same as those used to attain a balanced and representative test [cf. 9]. Similarity in difficulty is more a product of judgement than of explicit methodology. Still, it is easier to estimate the difficulty of multiple-choice questions than of short-answer questions.

11. The standard error of measurement is a measure of the precision of the test. It should not be used as the criterion by which to improve tests since it is only weakly related to test quality. The standard error is directly related to test length, and will decrease proportionately as the test is made longer.

12. Topics which seem to lend themselves only to error-prone test questions should be represented by more items on the test than topics for which it is easy to write good questions.
    12.1 The errors of measurement in a test score increase linearly as the number of test questions increases whereas the reliability of the score increases geometrically.
    12.2 The increase in reliability when more questions are included in a test is greatest when the initial reliability of the test is low. Marginal improvements in test reliability are smallest for tests with high reliabilities.
    12.3 More individual measurements are generally made in the area where each single measurement is least precise so that the total or average measurement in that area is acceptably reliable. If individual measurements are highly reliable, fewer need be made for an accurate estimate.

Chapter III Attitude measurement and survey bias

III. A. Introduction

The modular self-paced audio-tutorial system by which CEM 130 and 131 are taught is quite different from the typical university course. Because of this difference, the Student Instructional Rating System (SIRS) forms regularly distributed at Michigan State University for end-of-term course evaluations by students are inappropriate. A separate Course Evaluation Form was designed specifically for these courses. The one-page machine-scorable form consists of twenty-two items generally similar to those of the SIRS forms and with the distribution of item formats listed in Table 3.1.

Table 3.1 Distribution and description of the different item formats

Quantity	Item nos.	Description
two	1 - 2	Comparison of course with lecture
twelve	3 - 14	Statements with a Likert response scale*
three	15 - 17	Forced-choice from among ten course aspects
five	18 - 22	Multiple-choice items about varying topics

*Typical Likert scales range from strongly agree to strongly disagree.

Some who reviewed the results of the surveys thought that the definitely positive wording of many of the Likert items spuriously inflated the positive ratings generated by the survey. These criticisms stimulated a study of a comparison between this 'naturally biased' Likert format and an 'unbiased' format.

III.B. Literature

III.B.1. Attitude assessment techniques

Behavior is not necessarily the best indicator of underlying attitude. Because of time limitations or departmental requirements, a student may be constrained for or against chemistry courses independent of his attitude. As Guttman [74] says by way of anology:

> "It is conceivable that a person is 'against' a certain candidate but will vote for him because he is even more 'against' the opposing candidate, or he is 'against' a given proposition but will endorse it, if the only alternative is one he considers even worse; or he may be 'for' a candidate but even more so 'for' the opponent, or 'for' a proposition but even more so 'for' an alternative one."

Very few students who take these introductory chemistry courses are planning a major in chemistry, and the small number of students that decide for or against a major in chemistry during these courses is too small a sample to use as behavioral evidence of attitude change. The most straightforward method of determining the attitudes of students toward the course and its various aspects is to survey them in a questionnaire.

There are three major methods of constructing an attitude scale [75]: the method of equal-appearing intervals developed by L. L. Thurstone [76], the method of summated ratings described by Rensis Likert [77], and the method of scale analysis introduced by Louis Guttman [74]. In all three methods, a single attitude is measured by a collection of items.

Equal-appearing intervals. In this method a large collection of statements is ranked on a continuum from most favorable to least favorable by a judging group. The judges are instructed to order the statements with regard to the negative or positive nature of the statement notwithstanding their personal opinions. The statements are ordered along a nine or eleven point scale, and the median scale point for the group of judges is the scale value for a particular statement. The statements with the highest inter-judge agreement are selected for the final scale. An eleven (or nine) item questionnaire is constructed with one item at each scale point. A person responding to the survey is requested to indicate 'agree', 'disagree', or 'undecided' about each statement; the mean scale value of the statements with which he agrees is his attitude value.

Summated ratings. This method also begins with a large collection of statements on some attitude continuum. The person responding to an item indicates the extent of his agreement among several categories. The most common response key is a five-point scale such as

(1) strongly agree
(2) agree
(3) neutral
(4) disagree
(5) strongly disagree

The initial set of items is refined by standard techniques of item analysis and the most discriminating items are included in the attitude

scale so that about half of the items have 'strongly agree' as the favorable response and half have 'strongly disagree' as the favorable response. The responses are weighted according to their ordinal position in the response key from unfavorable to favorable, and the sum of the responses to all the items is the attitude value.

The Likert method is much more flexible than the Thurstone method since the number of responses and their wording can be varied. Research studies have shown the Likert method of scoring items to be consistently more reliable than the Thurstone method [78]. But there is still no external criterion which allows one to claim for either scale that a particular 'score' is the cutting point between favorable and unfavorable attitude. As Edwards [75] notes, there is no assurance that the 'neutral' response category is the true zero-point in the attituce continuum.

Scale analysis. Guttman's method does provide an approximate cutting point and is often able to estimate the absolute percent favorable attitude within the group. This method begins with a much smaller collection of items -- about as many as the number in the final survey. The items are tested for 'scalability'* by having a large group respond to them according to some Likert response scale. The final set of items includes only those which 'scale', although seldom if ever will a

*If the items are perfectly scaled, one who agreed with a mildly favorable statement would agree with all more favorable statements, and one who disagreed with a mildly unfavorable statement would disagree with all more unfavorable statements.

set of items scale perfectly. The person responding to the attitude scale indicates both his agreement or disagreement, and his 'intensity of feeling' about his response on a Likert-like scale. For the entire survey each person will have an agreement score and an intensity score. When these two scores are plotted against each other, a curve with a minimum intensity point at a particular agreement score will usually result. This agreement score is the cutting point between favorable and unfavorable attitude in the group surveyed. The technique falters when the general attitude is extremely skewed unless a large number of items near the intensity minimum are used.

Thus it is possible to measure the absolute attitude of a group toward a single thing. But a single course evaluation form elicits information about many student attitudes -- toward the course, toward the instructor, and about specific characteristics of each. It would require at least five or ten scalable items to determine each separate attitude and would increase the length of the current form from twenty-two to nearly one hundred items.

Since the development of the Thurstone and Likert techniques over forty years ago, no new techniques have received wide acceptance. As Shaw [79] observed in 1967,

> "...there seem to have been few major advances or breakthroughs in techniques of scale construction since the Thurstone and Likert methods were developed." "The overwhelming majority of scales has been developed by either the Thurstone or the Likert technique." "This is probably a result of the greater complexity of the newer procedures."

For practical reasons, variations of the Likert method of summated ratings are most often used for postcourse evaluation surveys. To simplify somewhat the ensuing discussion, the broad class of Likert variants may be divided into three types of item format. The first type is the classic Likert statement with responses on an agree-disagree scale. The second type is the evaluative item. Responses to the neutral statement in an evaluative item are along a judgemental scale such as superior-inferior, like-dislike, and such. The 'emotional content' of the item is in the response key and not in the item stem. A third type is patterned after the multiple-choice test question in which the answer choices are short descriptions of behavior or occurrence. This third type will be designated the 'descriptive' format. In all three types, the responses are ordered and assigned numerical values. The attitude score of an individual is the summation of his individual responses, and the group mean is the average response value for the survey or any single item.

III. B. 2. Effects of wording and structure in survey items

Even the smallest change in a survey item may greatly affect the pattern of response. The discussion of these changes and their effects is complicated by their intercorrelations, overlapping nomenclature, and lack of standardized taxonomy. The many specific changes investigated by different authors have been organized into four broad categories: (a) directionality, (b) fixed and variable responses,

(c) number of choices, and (d) specific wording. The literature in each of these areas will be briefly reviewed.

III. B. 2. a. Directionality in items or responses

Three related kinds of directionality may affect survey response. First, the direction of favorability in the attitude survey question may influence the response. The statement 'He is a good instructor' may not elicit a pattern of responses which is the mirror image of 'He is a poor instructor.' Second, the order of presentation in a compound sentence may influence the response. Again there may not be mirror-image response distributions when two sentences such as 'I prefer lecture to lab' and 'I prefer lab to lecture' are compared. Third, the order in which response categories are presented may affect responses.

III. B. 2. a. i. Direction of favorability in the item stem

Positively stated Likert items are widely believed to be subject to a response bias known as acquiescence. A student who 'acquiesces' would tend to agree with a positively stated evaluation item, thus falsely inflating the mean. Cronbach [80] and Guilford [81] both discuss acquiescence as a source of bias. Rorer [82] demurs, "...the importance of acquiescence is so widely accepted today that it has become necessary to demonstrate its nonexistence (rather than its existence, as would more appropriately be the case)." Based on a review of the literature, Rorer [83] concluded that in no instance was

acquiescence unequivocally demonstrated to be important.

Elliot [84] compared positively stated Likert items with the reverse or negative versions of the same items. The direction of the resulting bias depended on the content of the items. Sometimes acquiescence to the positive items produced the higher means; other times overcompensation on the negative items produced the higher means. The magnitude of the bias also depended on the aptitude level of the subjects responding to the survey.

This nonuniform occurrence of bias may be explained by the observation of DuBois [85] that there are two types of consistent responders. People who would respond 'neutral' or 'uncertain' when presented simultaneously with both polar alternatives may 'agree' to both or 'disagree' with both when presented separately. DuBois found about six percent of the responses in his study in the agree-agree category and about twenty-five percent in the disagree-disagree category. With any shift in attitude, the agree-agree person will show positive bias while the disagree-disagree person will show negative bias. DuBois did not describe the content of the scales used in his studies in the same terms as did Elliot, but the large disagree-disagree response pattern is consonant with Elliot's results on her opinion-neutral scale and the agree-agree response pattern is consistent with the results of her authoritarianism scale. From the meager examples given by the two authros this composite explanation seems reasonable.

The conclusion drawn from these studies of DuBois and Elliot is that the positive or negative bias present in Likert items depends on the specific content of the statement and on the aptitude of the student responding to the survey.

III. B. 2. a. ii. Order of presentation in a complex statement

Another type of item which shows the effects of directionality is the 'double-barreled' or 'rather-than' statement. Cantril [86] observed that the order in which alternatives in a compound sentence appear can sometimes influence the responses. "It appears that when the question is a fairly complicated one there is a tendency for the respondent to select the last, more easily remembered alternative." "In a number of other cases, interchanging the position of the alternatives failed to produce significant differences." This lack of consistency in the directionality effect may again be due to the influence of content demonstrated by Elliot.

III. B. 2. a. iii. Direction of favorability in response keys

The last directionality effect considered here is the order in which responses are presented in the key. Blumberg [87] found no effect. "In particular, it does not matter whether (a) the 'good' end of a graphic scale is located at the left, right, top, or bottom, (b) graphic scales or numerical ratings are used, (c) the order of presentation is one name at a time, one trait at a time, or a matrix with free choice

of order, and (d) subjects presented with the matrix choose to fill out one name at a time or one trait at a time."

In contrast, Mathews [88] found that the response was slightly biased toward the categories on the left. He suggested that reading habits may influence the response when the person responding to the item does not have a strong stable opinion. Mathews did not examine the effects of top-to-bottom ordering of the response key.

III.B.2.b. Fixed and variable response keys

Cronbach [89] considers the fixed-response format particularly prone to three kinds of bias: acquiescence, evasiveness, and extremism. Acquiescence has already been discussed in a previous section. Evasiveness, which some authors term 'central tendency' is the inclination to choose the neutral or middle category. Evasiveness leads to a narrowing of the distribution and a lessening of distinctions among what are rated. In contrast, extremism is the tendency to choose one or the other of the end categories. The result of extremism is to skew the distribution.

Cronbach suggests three ways to decrease bias. One possibility is to reduce the number of choices for the Likert item from five to two. With only two choices, there would be no evasiveness or extremism, but it seems likely that acquiescence might increase. Another possibility is to more objectively define the response categories for the Likert or evaluative items. This is akin to the multiple-choice format

wherein each response is a short description typical of that scale point. Still a third possibility is to use a forced-choice format (v. i.) in which all the choices appear equally favorable yet differ along some other dimension. Both the forced-choice and the descriptive multiple-choice are variable-response items, in contrast to the fixed-response Likert and evaluative items.

Guilford [81] agrees with Cronbach that fixed-response items are subject to various forms of bias. However, he continues, "...there are better ways of dealing with the same problems and ... the forced-choice device introduces some measurement problems that may be worse than those they were intended to correct." Guilford favors the descriptive multiple-choice format with responses given in operational terms. He counsels against the use of an evaluative format with such categories as 'excellent' or 'poor' because these terms are not well defined.

Both Cronbach and Guilford marshall attractive theoretical arguments to discourage the use of typical Likert and evaluative items. Yet neither cite specific experimental evidence that fixed-response items are more biased than variable-response items. Several studies of fixed- and variable-response formats have been made, but the results are mixed.

Sharon [90] constructed a forced-choice attitude survey which was quite resistent to bias. Each attitude item consisted of two characteristics known to correlate with teaching effectiveness, and two

favorable but uncorrelated characteristics. The students were requested to check the two most descriptive statements from each set of four. The total number of relevant statements checked by the student was the rating score for the instructor. The development of a good forced-choice survey involves the selection and scaling of a large number of statements for both general favorability and for correlation with teaching effectiveness.

The forced-choice format is really a hybrid between the Likert summated ratings and the Thurstone scaled statements. In addition to being as difficult to construct as any scaled survey, the forced-choice survey cannot be used for diagnostic purposes as can many simpler Likert scales. It does not indicate strengths and weaknesses, but only provides an overall rating of the instructor's effectiveness. It is mentioned here as a special variant of the descriptive format.

Smith [91] maintains that there is less acquiescence in the descriptive format than in the Likert fixed-response format. He also suggests that the social desireability of the multiple-choice responses be held constant to eliminate a favorability bias.

Stockford [92] compared a descriptive format with an evaluative format. He found that the evaluative scale was more positively biased and less reliable than the descriptive scale. Bryan [93] also compared an evaluative with a descriptive format. There was no significant difference between the two scales although the evaluative format often showed a small positive bias. The descriptions provided in the

multiple-choice items used by Bryan had labels corresponding with the evaluative response choices whereas the descriptive items used by Stockford were not thus labeled. This difference may explain the difference in leniency bias each author observed.

Showers [94] compared Likert, evaluative, and descriptive formats on the Student Instructional Rating System forms used at Michigan State University. Virtually all item means were markedly positive, but there were significant relative differences between forms. The evaluative format was consistently less biased than the Likert format. The descriptive format varied from most biased to least biased. The reliabilities of the three types of items were comparable. "The claimed superiority," concluded Showers, "of multiple choice over fixed alternative response cue formats in reducing bias was not substantiated by the results of this study of leniency."

However, neither the evaluative nor the descriptive items Showers used were rigorous examples. The response categories for the evaluative items are better characterized as descriptive-evaluative. Each evaluative label such as 'superior' or 'below average' was accompanied by a short defining phrase. The response categories for the descriptive items were not true descriptions but rather one or two word prompts. Further, only the middle and ends of the scale were defined by such terms; the second and fourth (of the five) choices were unlabeled. Because of these departures from the 'classic' patterns, Showers' results may not generally hold.

Elliot [84] compared variable-response descriptive items with fixed-response Likert items stated both positively and negatively. No single format was uniformly least or most biased. The effect of survey content on the biases of different formats may serve to reconcile some of the conflicting conclusions published by several authors. On the authroitarianism scale, the descriptive multiple-choice format was less biased at all aptitude levels than the traditional positively-stated Likert format. Since authors studying leniency bias most often used the authoritarianism or F-scale, the conclusion that fixed-response items are more biased than descriptive items is reasonable.

But the content of a course evaluation form most closely resembles Elliot's two 'personal' categories. With this survey content, the positive Likert items were always less biased than the descriptive items, and the negative Likert items were usually most biased. Thus the conclusion of Showers that descriptive items are not less biased than other formats does not necessarily contradict other cited literature when survey content is considered.

III. B. 2. c. Number of response choices

Although the most common number of response caterories is five, different attitude scales have been constructed with as many as a hundred or as few as two choices. Does increasing the number of categories increase the survey response reliability? This question

has been discussed theoretically and examined experimentally by many authors.

Cronbach [89] argues that increasing the number of choices beyond two only allows different biases to spuriously inflate the reliability. The inclusion of an 'undecided' category in an 'agree-disagree' key may lead to central-tendency or evasiveness bias. The differentiation between 'agree' and 'strongly agree' may allow extremism to bias the results. Cronbach asserts that while a greater number of response choices may increase the reliability, anything which allows response sets to bias the results can only lower the logical validity.

Remmers [49, 51] examined the reliability of attitude scores for two-, three-, five-, and seven-choice attitude items. The Spearman-Brown prophecy formula correctly predicted the reliabilities of all but the seven-choice items. Thus until seven choices was reached, the attitude item behaved like the multiple-choice test item and increased in reliability as the number of answer choices increased.

Matell [95] and Jacoby [96] studied the reliability of attitude items while varying the number of choices between two and nineteen. They found no relationship between the number of choices and either reliability or validity. They conclude that neither response bias nor reliability are influenced systematically by the number of scale steps and so concern about either should not influence the choice of how many response categories to use. One caveat to be added is that

when experimental groups are as small as those used by Matell and Jacoby -- twenty -- the reliability coefficients are quite imprecise; a true relationship may be lost in the noise of random error.

Bass [97] showed that as the number of categories increases, the overlap between them increases. Any fineness of distinction gained when the number of choices is incremented may be muddled by the overlap of adjacent categories. Overlap results from the range of meanings different words possess; it is discussed in the next section.

III. B. 2. d. Specific wording of statements or responses

Differently worded statements plumbing the same attitude will engender different proportions of agreement from the same individuals. The pattern of response will also change when the words in the answer key are altered even though the item stem remains unchanged. These variations in survey response are linked to the connotations of frequency or degree in a given word or sentence.

III. B. 2. d. i. Words of frequency and degree

Pratt and others [98, 99] distinguish between determinate numbers such as 'three' or 'seventeen' and indeterminate numbers such as 'few', 'often', 'seldom', or 'many.' The numerical meaning of an indeterminate number depends on the context.

Pepper [100] asked subjects to define the numerical occurrence of five frequency terms in several contexts as well as no context.

"The major finding was that the mean definition assigned to an expression embedded in a context shifted from its no-context definition toward the estimated frequency of the context event. For example... the mean definition of 'sometimes' [in the context of how often men found Miss Sweden attractive] was higher than the mean definition of 'very often' [in the context of earthquake occurrence]." "...it is easier for [subjects] to describe a low-frequency event with a high-frequency expression (California very often has earthquakes) than to describe a high-frequency event with a low frequency expression (Californians seldom eat dinner)." In light of these observations, it is better to use attitude statements which describe typical or easily estimated behavior than to state rare and uncommon conditions.

The Likert item should be stated in the words closest to the expected average circumstance. For example, the statement 'He is an excellent lecturer' is too extreme. A student who considers the instructor an average lecturer may disagree with this statement as much as a student who considers the instructor a poor one. The statement 'He is an average lecturer' overcorrects for extremism. The student who considers the lectures above average as well as the student who considers them below average must disagree with this statement. The recommended middle ground would be 'He is a good lecturer.' This statement is not so extreme that the bulk of the response is undifferentiated disagreement, nor is it centered such that disagreement could be interpreted as either positive or negative attitude.

Several other authors have also investigated the numerical values of such frequency terms, usually without context. Hakel [101], Simpson [102], Strahan [103], and Schriesheim [104] reported means, medians, and variabilities of many indeterminate numbers. The most narrowly defined words are those firmly anchored to a definite number -- i.e. 'always', 'never', and 'about as often as not' showed greatest stability.

Spector [105] scaled response words in three categories: evaluative, agreement, and frequency. These lists are presented in Figure 3.1. The author recommends that the words chosen for responses be equally spaced along these rankings.

III.B.2.d.ii. Variations in response keys

Dawes [106] compared five different rating scales as representational measures of height. Three correlated 0.94 with the objects rated while the fourth and fifth correlated 0.90 and 0.88 respectively. The scales with highest validity had an easily identified midpoint and a naturally stated symmetrical response key. One of the less valid scales was highly asymmetrical, while the other had too many intervals and no clear midpoint.

Dawes also noted that a bipolar response key can be biased. Generous-stingy is quite different from extravagent-thrifty. Both ends of the continuum should be stated in equally favorable terms so as not to court a social desireability bias.

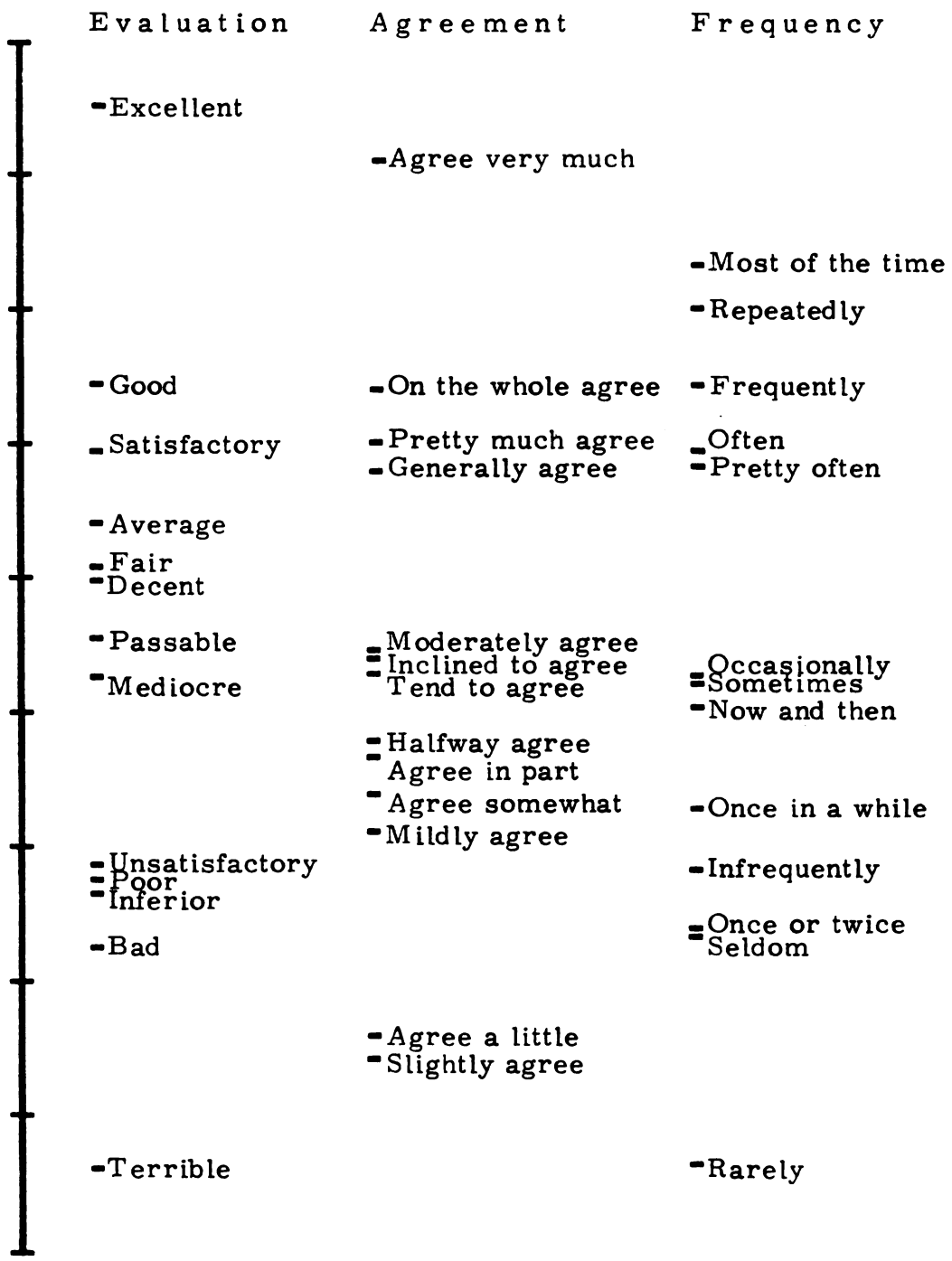


Figure 3.1 Relative scale positions of different response choices for three response continua, taken from Spector [105]. The scale is interval with arbitrary units.

Holdaway [107] examined five slightly different response keys for a set of Likert items. He found that the terms 'neutral' and 'undecided' have different meanings; a greater percentage of respondents chose 'neutral' and they appeared to come from both the agree and disagree sides of the scale. When the 'undecided' category was placed to the side of the scale, Holdaway observed a significant decline in the percentage who chose this response. He also observed a flattening of the distribution when Example A listed below was substituted for Example B.

Table 3.2 Examples of two response keys from Holdaway [107]

Example A	Example B
Agree	Strongly agree
Mildly agree	Agree
Undecided	Undecided
Mildly disagree	Disagree
Disagree	Strongly disagree

Edwards [108] and Guilford [109] caution that the 'undecided' or 'neutral' category, while it may be the midpoint of the scale, does not necessarily indicate a truly neutral attitude. If the Likert statement is favorable, a student may feel that to indicate neutrality he must disagree with it.

III.B.3. Summary of survey wording effects

No single item format is unequivocally least biased. The content of the survey question can determine which item format is likely to be least biased.

Descriptive multiple-choice items are subject to erratic spacing of response categories and unpredictable acquiescence to attractively phrased choices. In theory the descriptive format can show the least bias, but in practice it is extremely difficult to devise five different but equally unbiased choices for every item.

Fixed-response Likert and evaluative items may be subject to various response-set biases such as leniency, extremism, or evasiveness. Responses are not significantly affected by the order of choices in the answer key.

Rewriting a statement as its 'logical inverse' does not produce a mirror-image response pattern. Depending on question content and phrasing, the inverted form may be more or less lenient than the original statement.

Words or descriptions which allow the least personal interpretation yield the most reliable response keys. However, it is often difficult to design an answer key which is not only reliable and unambiguous but also equally spaced and not unduly lenient.

The number of choices in the answer key appears to be independent of either the reliability or validity of the scale. The most widely accepted designs use five response categories.

For the specific type of content found on course evaluations, the positively-stated Likert items are generally least lenient. The Likert item should be stated in words closest to the expected average circumstance. A too lenient Likert item can be made less biased by using a slightly less favorable wording in the stem statement. For example, if 'The audio tapes provided vibrant and comprehensive coverage of the topics' is too lenient, use instead 'The audio tapes provided good coverage of the topics.'

The middle category of a response key is not necessarily the truly neutral attitude. Some items may be worded such that a student may feel that to indicate neutrality he must disagree with a statement. The truly neutral midpoint is most likely to be found in examples of the evaluative format with a completely neutral and unbiased stem, and a simple symmetrical answer key. An example of such an item is:

> How much did you like this method of teaching a course compared to the traditional lecture-recitation method?
>
> (1) very much less
> (2) less
> (3) about the same
> (4) more
> (5) very much more

Both words in the antonymous pair 'more-less' are equally favorable, and the middle response clearly indicates a roughly equal liking for both teaching methods.

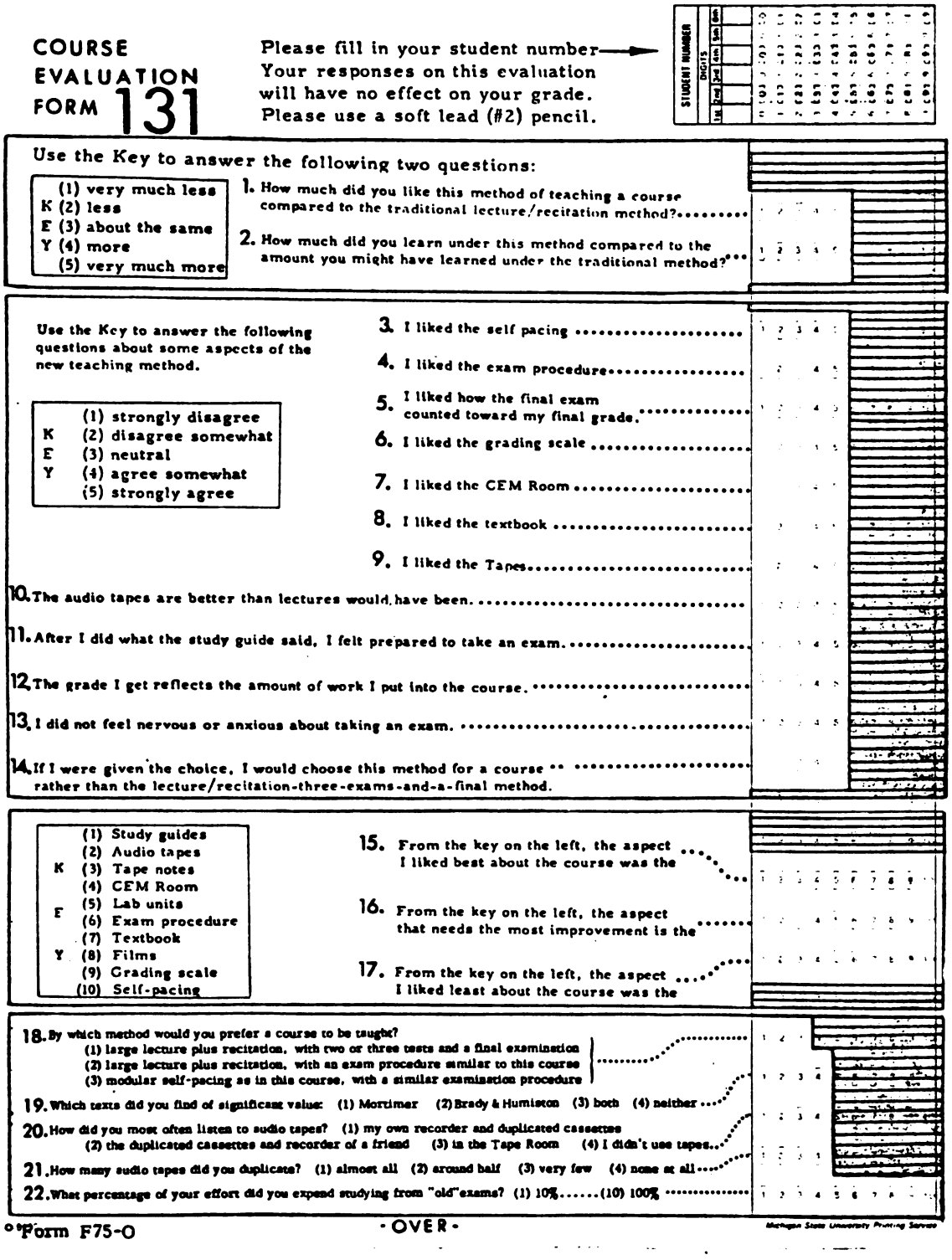
III. C. Experimental design

Two different postcourse evaluation forms were prepared for distribution to the students. In designing the 'new' form to be compared with the 'old' form, Items 1, 2, 11, 12, and 13 were not changed so as to serve as a baseline between forms. Items 3 through 9 were changed from the Likert format to an evaluative format -- a sample item and the two response keys are listed in Table 3.3. Items 10 and 14 are complex 'rather-than' statements and the logical inverse of the original statement appears on the 'new' form.

Table 3.3 Sample item in both Likert and evaluative formats

Likert		Evaluative	
I liked the self pacing...		Self pacing...	
(1)	strongly disagree	(1)	disliked very much
(2)	disagree somewhat	(2)	disliked somewhat
(3)	neutral	(3)	neutral
(4)	agree somewhat	(4)	liked somewhat
(5)	strongly agree	(5)	liked very much

The two forms were randomly collated with the final examination and distributed to the students during the final exam period. Response to the survey was voluntary; students were requested to fill out the surveys after completing their final exam and the forms were collected as they left the room. The forms are displayed in Figures 3.2 and 3.3.



COURSE EVALUATION FORM 131

Please fill in your student number →
Your responses on this evaluation will have no effect on your grade.
Please use a soft lead (#2) pencil.

STUDENT NUMBER
DIGITS
1st 2nd 3rd 4th

Use the Key to answer the following two questions:

K E Y
(1) very much less
(2) less
(3) about the same
(4) more
(5) very much more

1. How much did you like this method of teaching a course compared to the traditional lecture/recitation method?
2. How much did you learn under this method compared to the amount you might have learned under the traditional method?

Use the Key to answer the following questions about some aspects of the new teaching method.

K E Y
(1) strongly disagree
(2) disagree somewhat
(3) neutral
(4) agree somewhat
(5) strongly agree

3. I liked the self pacing
4. I liked the exam procedure
5. I liked how the final exam counted toward my final grade.
6. I liked the grading scale
7. I liked the CEM Room
8. I liked the textbook
9. I liked the Tapes
10. The audio tapes are better than lectures would have been.
11. After I did what the study guide said, I felt prepared to take an exam.
12. The grade I get reflects the amount of work I put into the course.
13. I did not feel nervous or anxious about taking an exam.
14. If I were given the choice, I would choose this method for a course rather than the lecture/recitation-three-exams-and-a-final method.

K E Y
(1) Study guides
(2) Audio tapes
(3) Tape notes
(4) CEM Room
(5) Lab units
(6) Exam procedure
(7) Textbook
(8) Films
(9) Grading scale
(10) Self-pacing

15. From the key on the left, the aspect I liked best about the course was the
16. From the key on the left, the aspect that needs the most improvement is the
17. From the key on the left, the aspect I liked least about the course was the

18. By which method would you prefer a course to be taught?
(1) large lecture plus recitation, with two or three tests and a final examination
(2) large lecture plus recitation, with an exam procedure similar to this course
(3) modular self-pacing as in this course, with a similar examination procedure
19. Which texts did you find of significant value: (1) Mortimer (2) Brady & Humiston (3) both (4) neither
20. How did you most often listen to audio tapes? (1) my own recorder and duplicated cassettes (2) the duplicated cassettes and recorder of a friend (3) in the Tape Room (4) I didn't use tapes.
21. How many audio tapes did you duplicate? (1) almost all (2) around half (3) very few (4) none at all
22. What percentage of your effort did you expend studying from "old" exams? (1) 10%.....(10) 100%

Form F75-O

· OVER ·

Michigan State University Printing Service

Figure 3.2 Course evaluation survey - 'old' form

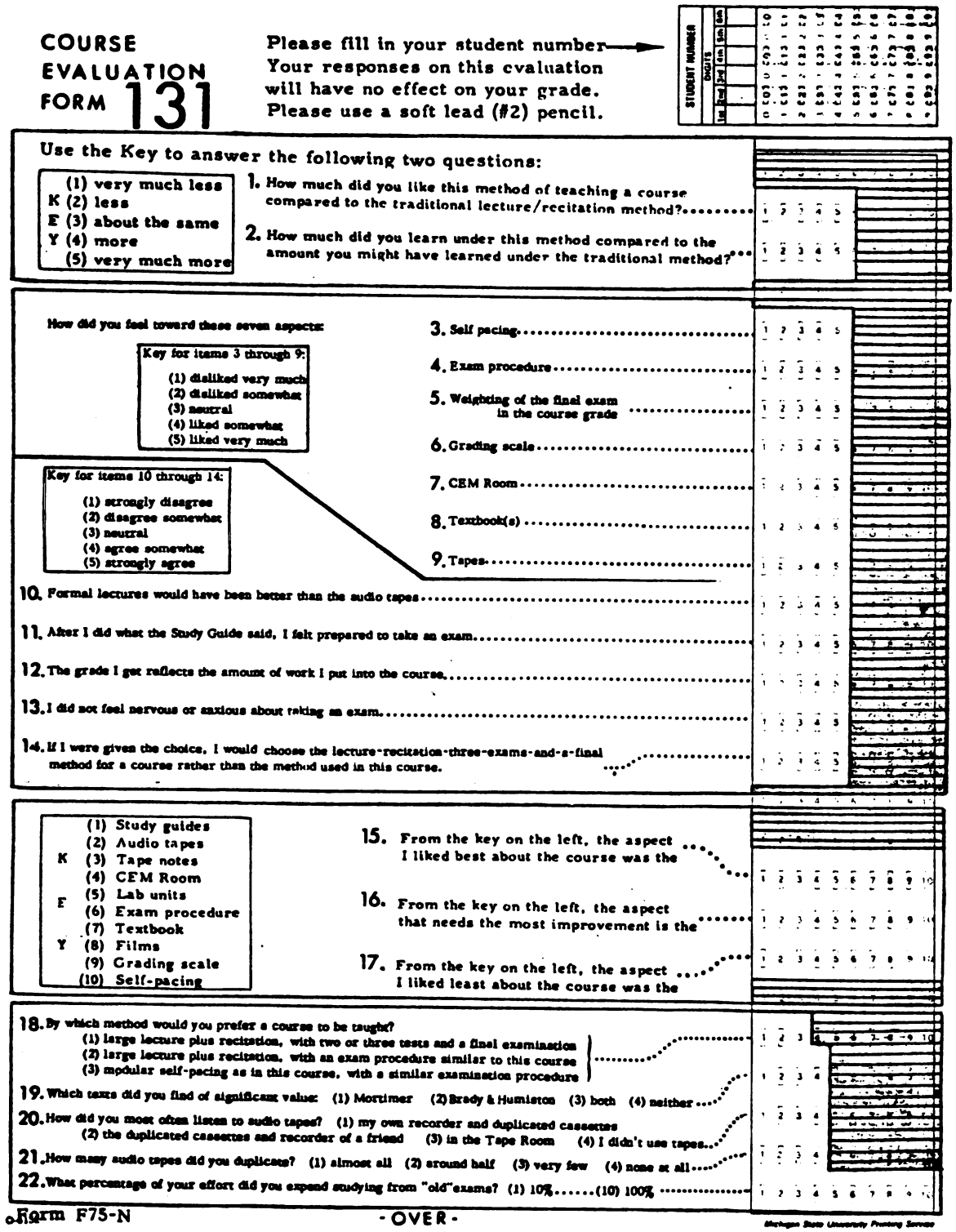


Figure 3.3 Course evaluation survey - 'new' form

III.D. Results and discussion

III.D.1. Tables of results and significance tests

An analysis of variance is not appropriate for these data. Items 1 and 2 on both forms were identical, yet in every experimental comparison the new-form mean is higher than the old-form mean. There is no explanation for this curious coincidence save chance. Since there is a consistent difference in favor of the new form on items which should have identical response means, the analysis of variance would likely yield a false positive. As a substitute significance testing procedure, multiple t-tests were used. This method is too liberal to verify the significance of specific results but it is adequate to identify any important trends in the data.

The differences in mean response for each item on the two forms and each t-test statistic are listed in Table 3.4. The new-form response values for Items 10 and 14 are inverted so that the sense of the numbers agrees with the original format. Two different levels of significance are indicated, based on two conceptions of these groups. If these results pertain only to the specific groups which were surveyed, the significance of a difference in response means depends on the percent survey return -- if the return is one hundred percent, any difference is significant. If however, the groups are treated as representative samples of an infinite population, then the appropriate critical t-value does not depend on percent return; the 'infinite' critical t-value for all tests is approximately 1.97. The 'finite' critical

t-values vary with percent survey return and are listed in the table for each group. Those differences which are significant under either consideration of these samples are indicated by a double asterisk and those differences significant only for finite populations are indicated by a single asterisk. Scattered significances will not be considered meaningful since the multiple t-testing procedure will indicate that about five percent ($\alpha = .05$) of the differences are significant by chance. Only if the same item shows a difference across most or all groups will the difference for that item be considered 'real.'

The ratio of variances on the old and new forms are listed in Table 3.5 along with the F-test statistics. If these variances are statistically equal, the ratios will lie between critical limits of 1.39 and 0.72. If a ratio is outside these limits then the variances are significantly different at the $\alpha = .05$ level. Only one of the eighty-four ratios is outside these limits, but since one would expect four or five significant differences by chance alone, this anomaly is not indicated.

Table 3.4 Differences in response means and t-test statistics for each item

Item	131W75		130W75		131S75		130F75		131W76		130W76	
	$\Delta\mu$	t	$\Delta\mu$	t	$\Delta\mu$	t	$\Delta\mu$	t	$\Delta\mu$	t	$\Delta\mu$	t
1	.02	0.195	.21	1.793*	.11	0.729	.07	0.894*	.15	1.636*	.04	0.350
2	.03	0.347	.26	2.634**	.10	0.305	.08	1.235*	.14	1.840*	.13	1.379*
3	.30	2.952**	.53	4.703**	.37	2.688**	.20	2.697**	.24	2.740**	.14	1.367*
4	.04	0.393	.17	1.546*	-.09	0.644	.04	0.607	.03	0.350	.06	0.674
5	-.17	1.833*	.13	1.240*	-.09	0.534	-.02	0.311	.14	1.723*	-.05	0.537
6	-.05	0.575	.20	2.026**	-.14	1.104	-.01	0.163	.11	1.506*	.01	0.118
7	.10	0.995	.13	1.204*	.12	0.836	.24	4.167**	.23	2.790**	-.08	0.790
8	-.04	0.448	.10	1.012*	.05	0.420	.19	2.946**	.21	2.760**	.18	1.506*
9	.07	0.747	.16	1.623*	.25	1.929**	-.08	1.144*	.48	5.842**	.03	0.305
10	.08	0.809	.16	1.339*	.38	2.688**	.08	1.025*	.19	2.169**	.07	0.632
11	.08	0.851	.04	0.370	.15	1.151	.17	2.562**	.16	1.909*	.09	0.962*
12	-.02	0.189	.17	1.394*	.05	0.330	-.01	0.132	.17	1.864*	.04	0.372
13	.02	0.201	.06	0.543	.20	1.420*	-.15	1.990**	-.08	0.960	-.09	0.874*
14	.55	4.974**	.75	6.028**	.26	1.666*	.36	4.428**	.39	2.564**	.40	3.476**
t_{finite}		1.300		0.900		1.300		0.840		1.120		0.860

Table 3.5 Significance tests for equality of variances

ITEM	CEM131W75 $\frac{s^2}{s^2}$*	F	CEM130W75 $\frac{s^2}{s^2}$	F	CEM131S75 $\frac{s^2}{s^2}$	F	CEM130F75 $\frac{s^2}{s^2}$	F	CEM131W76 $\frac{s^2}{s^2}$	F	CEM130W76 $\frac{s^2}{s^2}$	F
1	$\frac{1.765}{1.700}$	1.039	$\frac{1.804}{1.749}$	1.031	$\frac{1.723}{1.957}$	0.880	$\frac{1.852}{1.759}$	1.053	$\frac{1.724}{1.962}$	0.879	$\frac{1.836}{1.924}$	0.954
2	$\frac{1.203}{1.236}$	0.973	$\frac{1.287}{1.222}$	1.053	$\frac{1.410}{1.390}$	1.014	$\frac{1.213}{1.231}$	0.985	$\frac{1.169}{1.340}$	0.872	$\frac{1.192}{1.325}$	0.900
3	$\frac{1.742}{1.796}$	0.970	$\frac{1.997}{1.534}$	1.302	$\frac{1.878}{1.389}$	1.352	$\frac{1.882}{1.433}$	1.313	$\frac{1.742}{1.708}$	1.020	$\frac{1.609}{1.533}$	1.050
4	$\frac{1.829}{1.759}$	1.040	$\frac{1.815}{1.515}$	1.198	$\frac{1.782}{1.596}$	1.117	$\frac{1.396}{1.215}$	1.149	$\frac{1.605}{1.682}$	0.954	$\frac{1.203}{1.166}$	1.032
5	$\frac{1.528}{1.419}$	1.077	$\frac{1.573}{1.415}$	1.112	$\frac{3.108}{1.912}$	1.626	$\frac{1.295}{1.190}$	1.088	$\frac{1.554}{1.384}$	1.123	$\frac{1.402}{1.183}$	1.185
6	$\frac{1.344}{1.253}$	1.073	$\frac{1.439}{1.240}$	1.160	$\frac{1.489}{1.323}$	1.125	$\frac{1.299}{1.045}$	1.176	$\frac{1.196}{1.193}$	1.003	$\frac{1.101}{1.125}$	0.979
7	$\frac{1.720}{1.772}$	0.971	$\frac{1.683}{1.514}$	1.112	$\frac{1.879}{1.704}$	1.103	$\frac{1.337}{1.231}$	1.086	$\frac{1.401}{1.630}$	0.860	$\frac{1.530}{1.522}$	1.006
8	$\frac{1.453}{1.313}$	1.106	$\frac{1.429}{1.257}$	1.100	$\frac{1.327}{1.163}$	1.141	$\frac{1.321}{1.194}$	1.116	$\frac{1.282}{1.317}$	0.973	$\frac{1.337}{1.247}$	1.072
9	$\frac{1.481}{1.566}$	0.945	$\frac{1.393}{1.266}$	1.105	$\frac{1.341}{1.556}$	0.862	$\frac{1.462}{1.484}$	0.985	$\frac{1.472}{1.556}$	0.946	$\frac{1.481}{1.421}$	1.042
10	$\frac{1.674}{1.703}$	0.983	$\frac{1.974}{1.878}$	1.051	$\frac{1.596}{1.824}$	0.875	$\frac{1.843}{1.810}$	1.018	$\frac{1.607}{1.809}$	0.888	$\frac{1.738}{1.933}$	0.899
11	$\frac{1.450}{1.601}$	0.905	$\frac{1.661}{1.542}$	1.077	$\frac{1.569}{1.367}$	1.148	$\frac{1.354}{1.287}$	1.052	$\frac{1.324}{1.820}$	0.727	$\frac{1.313}{1.303}$	1.008
12	$\frac{1.830}{1.966}$	0.931	$\frac{1.993}{2.020}$	0.987	$\frac{1.954}{1.979}$	0.987	$\frac{1.674}{1.749}$	0.957	$\frac{1.848}{1.860}$	0.994	$\frac{1.714}{1.745}$	0.982
13	$\frac{1.635}{1.762}$	0.928	$\frac{1.589}{1.709}$	0.930	$\frac{1.628}{1.776}$	0.917	$\frac{1.715}{1.680}$	1.021	$\frac{1.601}{1.480}$	1.082	$\frac{1.602}{1.549}$	1.034
14	$\frac{1.963}{2.154}$	0.911	$\frac{2.252}{1.990}$	1.132	$\frac{1.994}{2.187}$	0.912	$\frac{2.122}{1.844}$	1.151	$\frac{1.847}{2.151}$	0.859	$\frac{2.078}{1.899}$	1.100

*Old form in numerator, new form in denominator

III.D.2. Differences between the two forms

The only consistent significant differences between the two forms are on Items 3 and 14. These two items had significantly higher means on the 'new' evaluative form for five experimental groups, and the 'new' means were higher but not statistically significantly so for a sixth group.

There were forty-two item by group comparisons of Likert and evaluative formats. Thirty-one times the Likert format was less biased and eleven times the evaluative format was less biased, but all twelve of the significant differences indicated the Likert format to be less biased.

The inversion of Items 10 and 14 clearly did not yield mirror-image responses from the students. All twelve item by group comparisons -- seven of them significant -- showed less leniency bias in the original construction.

However, the five common items between surveys also showed consistent bias. Twenty-five of the thirty comparisons were less biased (only two significantly so) on the 'old' form even though the items were identical on both forms. This happenstance damps any extravant claims about the overall leniency bias of the evaluative format as compared with the Likert format. Still, the original formats for Items 3 and 14 show less positive bias even after adjusting for the baseline differences between surveys.

III.D.3. Discussion of results

The results of this study indicate that the Likert format is not more biased than the 'unbiased' evaluative. Only one Likert item differs significantly from its evaluative counterpart and the Likert item is less biased, not more. The only work which might be directly compared with this is that done by Showers [94] on Student Instructional Rating System forms. Her conclusion was that the Likert format is more biased than the evaluative format. Four possible resolutions of this apparent conflict will now be presented. Any or all of these four suggested explanations may contribute to the observed differences between Showers' results and these.

First, the content of the surveys in the two studies may be different enough to bar comparison of the results. As Elliot [84] demonstrated, different areas of content change the biases in different formats. The surveys used by Showers ask the students to rate their instructor and his course on several characteristics, most of which make specific reference to the person of the instructor. Taylor [110] and Stockford [92] point out that when persons are rated, the ratings are more linient when the rated person will see the results, or when the ratings will be used to evaluate the person. In contrast to the many 'personal' SIRS items, none of the items on the course evaluation survey used in this study ask the student to directly evaluate an instructor in a classroom setting. This difference in survey content may influence the responses significantly.

Second, Showers asked for qualitative judgements of primarily external things along a scale of superior to inferior. When a student is asked to judge something, it must not only be easily observable but also clearly anchored to a standard of comparison. An instructor's enthusiasm, concern, or knoledgeability are less 'observable' than a student's own liking for the audio tapes, grading scale, or textbook. Students may be inclined to give the instructor the benefit of the doubt concerning that which they cannot readily observe.

Third, the formats compared by Showers may not be the same as the formats compared in this study. The difference between Showers' Likert and evaluative items is deeper than simply a change in format. The nature of this difference is clarified by the examples in Table 3.5.

Table 3.5 Examples of different formats for the same survey item

Evaluative(anchored)	Likert(unanchored)	Likert(anchored)
Instructor's enthusiasm was....	Instructor was enthusiastic.	Instructor's enthusiasm was above average.
Superior	Strongly agree	Strongly agree
Above average	Agree	Agree
Average	Neither A nor D	Neither A nor D
Below average	Disagree	Disagree
Inferior	Strongly disagree	Strongly disagree

Showers compared anchored evaluative items with unanchored Likert items. The response to an unanchored Likert item can be quite different from the response to an anchored item. If the students

perceive all instructors as enthusiastic, then an enthusiastic instructor is merely average. Here the observed difference between the mean responses on two different survey formats is caused not just by rewording the same question but by asking a different question. The anchored item asks for a judgement relative to an assumed norm, the unanchored item asks for an absolute judgement. To the extent that this relative-absolute dichotomy is present, the two questions will have the same responses only by accident. When the unanchored Likert item has a higher mean than the evaluative item it may only mean that the occurrence of the quality in question is widespread. A truer determination of the effect of the format would be gained by comparing the anchored Likert item with the anchored evaluative item, as was done in this study, or by comparing the unanchored Likert item with an unanchored evaluative format. (An unanchored evaluative format avoids any reference to the average or typical -- e.g., five response cues might be: excellent, good, passable, poor, and terrible.)

Fourth, it is possible for surveys to show a positive bias in one situation and a negative bias in another situation. This reversal can only happen when the response intervals on one survey are smaller than those on another survey. By analogy, consider the Fahrenheit and Celsius scales. The Fahrenheit temperature is a higher number than the Celsius temperature whenever temperatures are above -40^{o}. But the Celsius temperature is a higher number than the Fahrenheit when temperatures are below -40^{o}. In a sense, the Fahrenheit scale is

more lenient in the first case and the Celsius scale is more lenient in the second case. In contrast, the Kelvin scale is always 'more lenient' than the Celsius scale and gives a more positive number across all range of temperatures. Thus if the Likert and evaluative surveys have equal scale intervals, the observed response mean of the more lenient scale will be consistently biased in the same direction. If the Likert scale has narrower response intervals, then the direction of the observed bias will depend on the response means. If the observed means are high, a Likert scale may show positive bias relative to an evaluative scale, whereas if the observed means are low, the opposite may be seen. The response means for the surveys used here and by Showers are listed below in Table 3.7.

Table 3.7 Distribution of survey response means

Range	Showers*	Kales**
4.00 - 5.00	21.6	0.0
3.50 - 3.99	60.8	21.4
3.00 - 3.49	17.6	25.0
2.50 - 2.99	0.0	50.0
2.00 - 2.49	0.0	3.6

*Percent of the fifty-one means in each range
**Percent of the twenty-eight means in each range

The distributions of item means are quite different in the two groups of data. Therefore, in at least one of these studies, it may be that the scale intervals on one of the surveys are narrower than those on the other. The variance of the items in this study do not differ significantly, thus supporting the equal-interval assumption for the Likert and evaluative formats used here. Showers did not explicitly test for equality in her study, hence the validity of the equal-interval assumption for her data remains unsubstantiated.*

Although any of the four above explanations singly or in combination may account for the different conclusion of Showers, it seems most likely that her comparison of the unanchored Likert scale with the anchored evaluative scale is the principal determinant of her results. In sum, over the six groups and more than four thousand students in this study, the Likert format as originally written is no more lenient and sometimes significantly less lenient than the 'naturally unbiased' evaluative format.

The second significant bias observed in these results is that of the inverted items. There is a consistent and often dramatic difference between what appear to be mirror-image statements. The original and

*The multivariate analysis of variance technique which Showers used is not sensitive to violation of the assumption of equal variances among the groups. However, if a trend exists in the data whereby one of the surveys tends to have a larger variance, then the differences in the means may be due to the above explanation. The conclusion would have to be qualified by stating "In this region of the scale....format L is more lenient than format E."

inverted forms of Item 14 are presented in Table 3.8. More students tend to disagree with the inverted statement than agree with the original statement. The responses to either form of Item 14 also tend to be more polarized than responses to other survey questions, as can be seen in the larger variances for this item.

Table 3.8 A survey question involving a 'rather-than' comparison

Format	Statement
Original item 14	If I were given the choice, I would choose this method for a course rather than the lecture-recitation-three-exams-and-a-final method.
Inverted item 14	If I were given the choice, I would choose the lecture-recitation-three-exams-and-a-final method for a course rather than the method used in this course.

Guilford [109] postulates that "...natural attitude attempts to dichotomize experience, avoiding the indifferent and the mediocre in affective reactions." Item 14 presents an either-or choice to the student and the responses are perhaps therefore more readily polarized than for other questions.

Hoveland and others [111, 112] state that subjects show a tendency to extremeness in rejecting statements with which they disagree. Gordon [113] found that negative extremeness was more pronounced than positive extremeness in several attitude surveys he investigated. Warr [114] suggests that the extremeness tendency is

a function of the involvement associated with the task. If students perceive the matter as important and involving, they tend to respond more extremely than otherwise.

Survey question No. 10 does not show such differences between the original and inverted wording as does No. 14. The content of No. 10 asks for a more ambiguous judgement of whether the audio tapes are bettwe than lectures would have been. As a result, the group variances and differences in means tend to be lower than those for question No. 14.

III.D.4. Conclusions

On the whole, the Likert items with "I liked such and such" stems were slightly less biased than the evaluative items with "How did you feel about such and such" stems. Only one of the seven Likert survey items showed consistent significant bias compared with the evaluative format, and the bias was in the negative direction -- i.e., the Likert item was less lenient than the evaluative item.

Of the two items which were inverted and compared, only one displayed a significant lack of mirror-image response. Item 14 concerned the concrete action of choosing in contrast with the quality judgement requested in Item 10. The easier it is in the student's own mind to make a distinction, the more likely his response will be extreme. It is also more likely for disagreement to be extreme than for agreement to be so. The student may find it easier to assess what choice to make rather than whether one thing is 'better' than another.

III. E. Correction for the observed bias

Survey Items 3 and 14 are the only two questions which showed large consistent differences between the original format and the experimental format. A factor analysis of the evaluation survey grouped Items 1, 2, 3, 14, and most of 10 into the same factor. Since Items 1 and 2 are the same on both forms, the difference between the new and old forms on these two items is a difference of group and not of format. This baseline difference is subtracted from the differences between the old and new forms of Items 3 and 14 so that the correction of the experimental format is based on only the net difference between the new and old formats.

The differences between the response means on the old and new survey formats are listed in Table 3.9 on the next page. Based on the average biases for the new formats of Items 3 and 14 two simplified corrections were made on the data. The first correction is of cell populations in the frequency count of responses to these items. The response mean on the experimental format is 'too high' by a certain fraction for these two items. To lower the response mean, the cell populations were each adjusted by shifting a fraction of the number of responses in each category to the next lower category. The fraction shifted downward for Item 3 is one-fifth and for Item 14 it is one-third. An example of this correction to cell frequencies is provided in Table 3.10. These corrected response frequencies for the experimental format are added to the response frequencies for the original format,

Table 3.9 Percentage differences between survey means for the original and experimental formats*

Term	Items 1 and 2 mean difference	Item 3		Item 14	
		raw	net	raw	net
131W75	2.5	30.0	27.5	55.0	52.5
130W75	23.5	53.0	29.5	75.0	51.5
131S75	10.5	37.0	26.5	26.0	15.5
130F75	7.5	20.0	12.5	36.0	28.5
131W76	14.5	24.0	9.5	39.0	24.5
130W76	8.5	14.0	5.5	40.0	31.5
mean			18.5		34.0
median			19.5		30.0

*All differences are positive since the experimental mean was always higher than the original mean.

Table 3.10 Example of a cell population correction - 130F75 No. 3

Identification	Response cell				
	1	2	3	4	5
Experimental cell frequencies	36	74	99	205	190
Number shifted downward	0	14	19	41	38
Sample calculations	(36 - 0) + 14		(205 - 41) + 38		
Corrected cell frequencies	50	79	121	202	152
Likert format cell frequencies	81	67	83	190	180
Cumulative cell frequencies	131	146	204	392	332

and the overall means were calculated. These corrected overall means are the basis for summary data for the course evaluation surveys presented in Chapter IV and the appendix.

The second correction is to each each response to Items 3 and 14 in the experimental format. The numerical values attached to the five response categories are decreased by an amount roughly equal to the net difference between the original and experimental formats. Every response to Item 3 was decreased by 0.2 so that a response of 4 is given the value 3.8, and so on. Responses to Item 14 were each decreased by 0.3. These individually corrected responses are used in the calculation of correlation coefficients between survey response and course grade reported in Chapter IV.

The correction formulas are only approximations; they mitigate but do not eliminate the biases between the experimental and original item formats for Items 3 and 14. Responses to all other survey questions are uncorrected; the grand means are based on the simple summing of response frequencies.

III. F. Summary for attitude survey bias

The Likert items which appeared on the original design of the course evaluation form are not significantly biased compared with the 'unbiased' evaluative format save for one item. The use of the Likert format allows the same answer key to be used for Items 3 through 14, and for this reason it will be retained in future editions of the survey.

The complex questions posed in Items 10 and 14 elicit different patterns of response when the sense of the sentence is inverted. The original construction is less positively biased than the inverted format and as such will be retained as the preferred form.

So that survey response from the alternate format may be compared directly with the original format from the same and other terms, the means and responses to Items 3 and 14 were corrected for positive bias on the experimental format. Responses to Items 3 and 14 were decreased approximately one-fifth and one-third of a scale unit respectively. No other change in responses to other questions was made. The results of the evaluation survey and the patterns of response are summarized and discussed in Chapter IV.

Chapter IV Student achievement and attitude

IV.A. What are students expected to know?

An inventory of course material can have several forms, each best suited to a different purpose. The general goals of CEM 130/131 stated in a course description are the name, size, and location of the 'country' of these courses in the world of academic subjects. The syllabus for these courses is a partitioning of this cognitive area into divisions and subdivisions which maps their interrelationships and relative importance. Study guide units are a division of the syllabus into equal-density parcels equivalent in these courses to one day's work. The smallest division of the material covered by CEM 130 and 131 is into testable concepts. The last three of these cartograms -- syllabus, study guides, and testable concepts -- will be discussed in the next sections.

IV.A.1. Syllabus of topics

The course syllabus for CEM 130 and 131 is a key-word outline of what students are expected to know when they finish these courses. The thirteen topics in the syllabus cover two-thirds of a one-year sequence in chemical principles. It was completely rewritten when these courses were redesigned, but is closely derived from the previous syllabus.

The syllabus provides a sequential map of what is to be taught and learned, but does not indicate how much time might be spent in

the presentation of a topic. The primary use of the syllabus is to organize the subject matter into its major divisions and to describe the prerequisite and subordinate relationships between topics and subtopics. A separate division of the syllabus into sections of equal time and effort is denoted by the succession of units. The complete syllabus for these courses is presented in Table A.1 in Appendix A.

IV.A.2. Study guide units

The sequence of 483 line-entries in the combined syllabus for CEM 130/131 is divided into sixty-three units of work -- some units span as many as 25 or 30 items, others as few as three. The study guide for each unit provides a statement of the general objectives of the unit. These objectives are a more detailed statement of what the student is expected to know after working through the study guide. A list of objectives for each unit is given in Table A.2 in Appendix A.

When unit objectives are stated in behavioral terms (for example, CEM 130 Units 31-32, p. 284) three important uses may be made of them. First, the instructor clarifies for himself exactly what he expects of the student and thus exactly what to teach. Second, the student is told the specific skills or knowledge he is expected to master and demonstrate on an examination. Third, test item selection and writing is greatly simplified since a behavioral objective is the generalized wording of a possible test question.

IV. B. Measuring knowledge of chemistry

Whether a student has mastered what he was taught is estimated bu giving him a test and recording his performance on it. In these courses examinations are required approximately every five to seven units*; at the end of the term a cumulative final examination is given. If only one form of an exam need be constructed, the instructor may go through the syllabus and study guides for the units to be covered and select directly the question ideas for his test. However, students in these courses are allowed to take tests at different times and to repeat their tests during the intervals** tests are offered. Thus many different forms of each exam are needed, and a more organized approach to test construction is required.

A test is not a unique set of questions but only one of many possible samples from a universe of all possible questions. If this pool of items already exists, test construction is merely a sampling process. The first step in constructing a pool of test questions from which the various exams will be sampled is to specify all the testable concepts in each of the units of material.

*There are six exams in CEM 130 and five in CEM 131. Some previous terms required as many as nine exams. Until 1976 the fourth exam in CEM 130 was a cumulative exam over the first twenty units.

**Originally each examination was offered for about fourteen class days; currently the exam 'windows' are seven or eight class days.

IV.B.1. Testable concepts from each unit of material

For each unit in the syllabus, as many different types of questions were written as could be conceived. The questions were grouped according to similarity of underlying concept. Some units contain many distinct testable concepts while others languish with one or two ill-defined categories. It was particularly difficult to produce many testable concepts for units of an introductory descriptive nature. In contrast, units involving problems and calculations may quite easily have ten clearly defined testable concepts. The list of testable concepts for each unit in the syllabus is presented in Table A.3 in Appendix A.

Once a concept has been defined, either by example or by a brief description, several questions are written for that concept. Testable concepts should be narrowly enough defined so that all questions to be included under that heading can be judged equivalent by inspection. Again, not only are concepts involving calculations easier to define, it is also easier to write numerous similar problems for such a concept. In CEM 130, which covers much descriptive chemistry, there are an average of only nine questions per concept, whereas in CEM 131, which includes calculations involving gas laws, equilibrium, and solution chemistry, there are about fifteen questions per concept. The item pool (as of 1976) contained 1240 questions in CEM 130 and 1450 in CEM 131. The size of the question file is in a continuous state of flux as questions are deleted, changed, or added to the pool.

Increasing the quantity and quality of the test items in the exam file is a continuous process. The initial struggle was to achieve a number of questions large enough so that it is unlikely any question would be used more than once. For six exams of fifteen items, each given in twenty different forms, 1800 items are necessary; for five exams, 1500 items are sufficient. Since the file for CEM 130 fell short of this 'working minimum', test questions did appear on more than one form of an exam.* If more exams are planned in the course, or if each exam requires more than twenty forms, then either the size of the file would need to be increased, or more questions would encore on the tests. Conversely, if fewer forms or fewer exams were used, then a smaller pool of test items would suffice. Currently (1977) the question file contains over 2000 items for each course, and no item appears more than once each term.

IV.B.2. Computer-managed exam file

Test questions are stored in computer memory with an identification number denoting questions for a testable concept, concepts within a unit, and units in the pool. A computer program r randomly selects test questions from this pool according to specifications determined by the instructor. The hierarchy of numbers

*The assumption was made that twenty different forms of each exam are prepared. If only fourteen forms are prepared, virtually no item need appear more than once.

which identify each question allows the instructor to specify any level of stratification from simple random sampling of the entire pool to selection of specific items. This is accomplished as follows.

Generation of an exam begins with specifying the examination composition index (ECI). A range of units and concepts is stated for each of the fifteen items on the test; within this range the computer will sample randomly one question without regard for unit-concept boundaries. If all fifteen ranges are identical, then the selection process is simple random sampling from among all questions in the specified range. If each range is different, then selection is random within the fifteen defined strata. When the initial and terminal identifiers are identical, then the question denoted by those numbers is guaranteed of selection. There are no restrictions on how large or small the range of questions specified in the ECI must be, and ranges may vary in size for a single exam from one item to the entire pool. Within the range specified by the ECI, no question will be chosen twice in one term until all questions have been used once. Whether questions are repeated on different forms of an exam is thus a function of the number of forms generated and the number of items within the given range.

The number of different question ranges in the ECI specification is the degree of stratification in the sampling process. The practice in these courses is to define fifteen separate strata within which questions will be randomly selected for each exam. This maximum

level of stratification is required by the nature of the material. As demonstrated in Chapter II, there is considerably more variation in content within a single exam than between two parallel forms. For heterogeneous subject matter, stratified random sampling of test questions ensures a more representative estimate of what students know than would simple random sampling.

The process described above produces a set of measuring instruments any one of which may be used to estimate a student's knowledge of chemistry. Although exams are not perfectly equivalent (as experimentally determined in Chapter II), they are all of equal validity in measuring knowledge of chemistry.

IV. C. What did the students learn?

The question 'What did the students learn?' is not answered by simply stating the topics which were presented. Only when the course is taught according to a mastery model can it be said that a student who finished the course learned all or nearly all of the material.

IV. C. 1. Overall achievement

These two courses denote the amount of chemistry learned by an eight-step grade scale collapsed from scores on the five or six proficiency examinations and a final exam. Students are allowed to retake examinations until they have achieved a grade with which they are satisfied, or until the closing date for that exam. Thus for each student three scores will exist by which their achievement may be measured: mean score on all tries of an exam, last-try scores for the exams, and final grade in the course. Each of these three measures provides somewhat different information about the 'true' level of achievement. Course means for these three quantities are presented for each term in Figure 4.1 Complete summary statistics for each term are presented in Appendix E.

Last-tries mean percent. The average score for a student of only the last time he attempted each exam is not necessarily the highest possible average. Frequently a student will attempt to raise his grade by retaking an exam only to score lower on the retake. If the closing date for an exam prevents that student from subsequently

raising his score back to its former level, then his last-try score is not his highest. If the mean score upon which grades are based were to be calculated from the highest scores of each student on each exam, the resulting means would likely be significantly higher. Yet the last-tries mean does serve as an upper bound to any estimate of overall achievement. It is unlikely that the mean of an entire class is ever underestimated by the last-tries mean percent.

All-tries mean percent. The average score a student achieves on all attempts he makes of an exam serves as an approximate lower bound to his actual knowledge. Most students study and learn more by the time they take their 'last try' on an exam, so that earlier low scores do not reflect how much they know when they are finished studying a particular section. Some students (though not many) simply take an exam before studying any of the material in order that they might get an idea of what the instructor will ask on the examination.

Final grade mean. The last-tries mean for each student is converted into a course grade according to a preset scale. The grade is sometimes modified by the term-end cumulative examination in a manner described in Section I.B.2.d. Average course grades reflect changes which were made in the grading scale; because of these changes grade means do not always show the same pattern as do the mean percent scores. In fall of 1974 the grading scale was made significantly more difficult, and in fall of 1975 it was made slightly easier. Both of these changes greatly affected the mean final course grade.

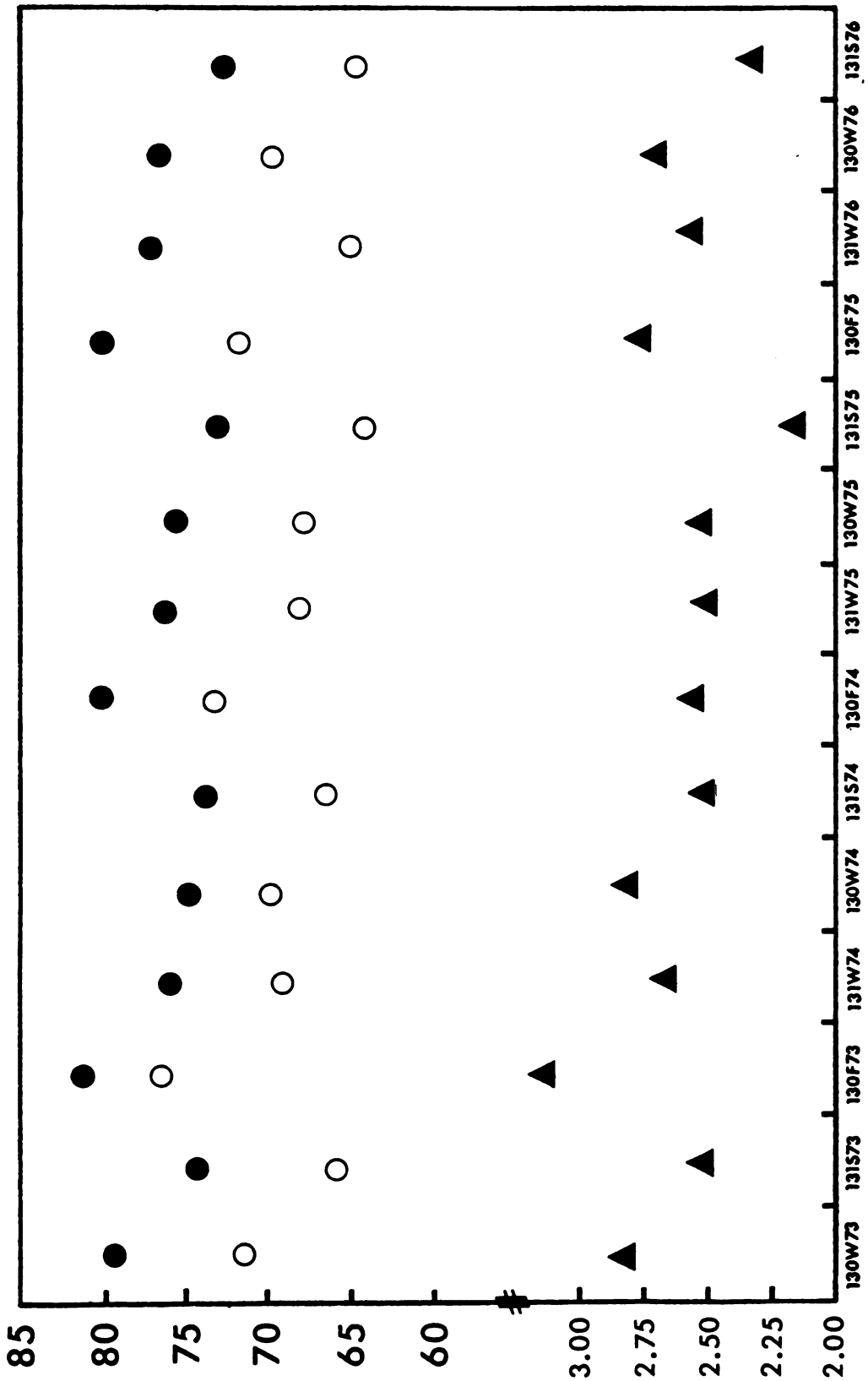


Figure 4.1 Last-tries mean percent [●], all-tries mean percent [○], final grade mean [▲]

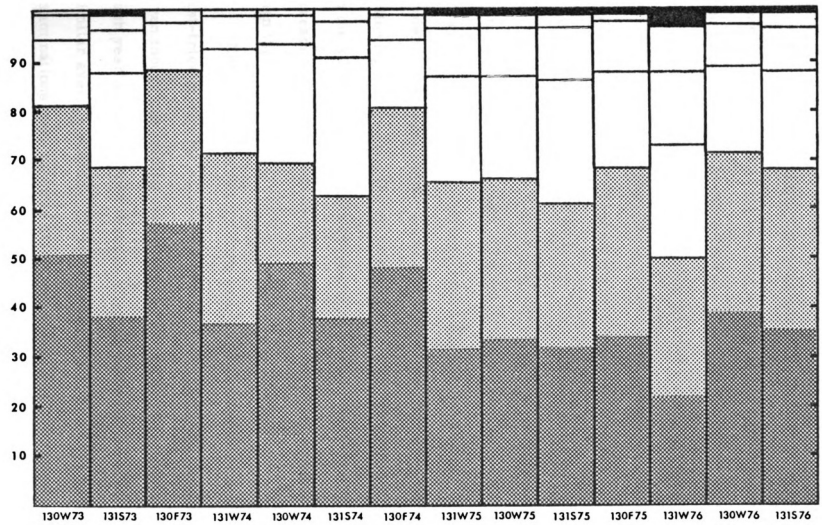


Figure 4.2 Average percent of the enrollment that attempted each of the exams: once - dark screen; twice - light screen; white blocks - three, four, and five times; black - six or more times.

IV. C. 2. Patterns of achievement

There are two distinct dimensions to the pattern of student achievement in CEM 130/131. One pattern is the periodicity of overall achievement in these courses within the academic year. The other is the variation in achievement from one exam to the next within the same course. Both patterns appear to be remarkably persistent, thus they bespeak some generic characteristic of either the course material, the students, or both.

As can be seen in Figure 4.1, the term averages proceed through a similar pattern each year. The mean for a Fall Term CEM 130 is always highest and for a Spring Term CEM 131 the lowest. Most students who enroll in one term of CEM 130 also enroll in CEM 131 the following term -- and the class average is always significantly lower in the second course of this two-term sequence. Also, the 'main stream' sequence which begins in the fall always has a higher mean than the 'trailer' sequence beginning in the winter.

Not only is the pattern from year to year similar, but also the last-tries means for successive years of each term are nearly identical. Even though the grading scales and the exam file underwent changes, each year's students took tests as often as necessary to achieve similar average scores. (This variation in the frequency with which examinations were taken is displayed in Figure 4.2.) It seems reasonable that the large enrollments in these courses represent groups with relatively stable capabilities, so that one Fall Term's

class is much like the next. One would then expect that each year's class would learn about as much chemistry as the last, and that scores on equivalent tests would be the same. Since the tests are not perfectly parallel, some deviations will occur -- when (as was the case) unrealistically easy items in the exam file are replaced by better (and more difficult) questions, then the expected scores would be lower even if the absolute amount which the students learned were the same. The repeatability of exams counters this effect by permitting students to retake exams and raise their scores. Like a self-correcting mechanism, if the tests are easier students retake them less frequently; if they are harder students take them as frequently as five or six times.

The second overall relationship in course means is that between CEM 130 and CEM 131. Students average four percentage points lower in CEM 131 even though they attempt each exam more often. Clearly students find CEM 131 more difficult than CEM 130. CEM 131 requires greater proficiency in algebra and less rote memorization than does CEM 130. The second course also builds on what was learned in the first course, and if a student forgot or never learned some of the material in CEM 130 later needed in CEM 131, his progress is slowed if not halted altogether.

Within each course there is also a repeating pattern in exam means. This variation in exam means indicates the relative difficulty of the different topics covered within each course. The mean percents for all tries and for last tries of each exam in these two courses are

displayed in Figure 4.3. Since the number of exams given in a course has changed several times since this method was introduced, it is difficult to compare two terms with different numbers of exams, but the periodicity of means for terms with the same number of exams is readily apparent.

IV. C. 3. Accuracy of the final grade

The accuracy of the grade given a student in CEM 130/131 is not a mathematically calculable statistic. Unlike the precision of a test score, the accuracy cannot be experimentally determined, but is based on several considerations of a rational rather than mathematical nature. The four variables considered in judging the accuracy of the grades given in these courses are: the reliability or precision of the term average, the representativeness of the tests, the grading scale, and the effect of the final exam. Each of these will be discussed with respect to the likelihood of assigning the wrong grade to a student.

Reliability of the term average. The reliability of individual fifteen-item tests typically range between 0.69 and 0.75. The reliability of the term average based on five or six tests is significantly higher than the individual test reliabilities. If it is assumed that the reliability of a single test is within the above stated range, then the cumulative reliability of a student's scores will lie between 0.87 and 0.91.*

*Calculation of the cumulative reliability of several tests is estimated according to the values displayed in Table C.2 in Appendix C.

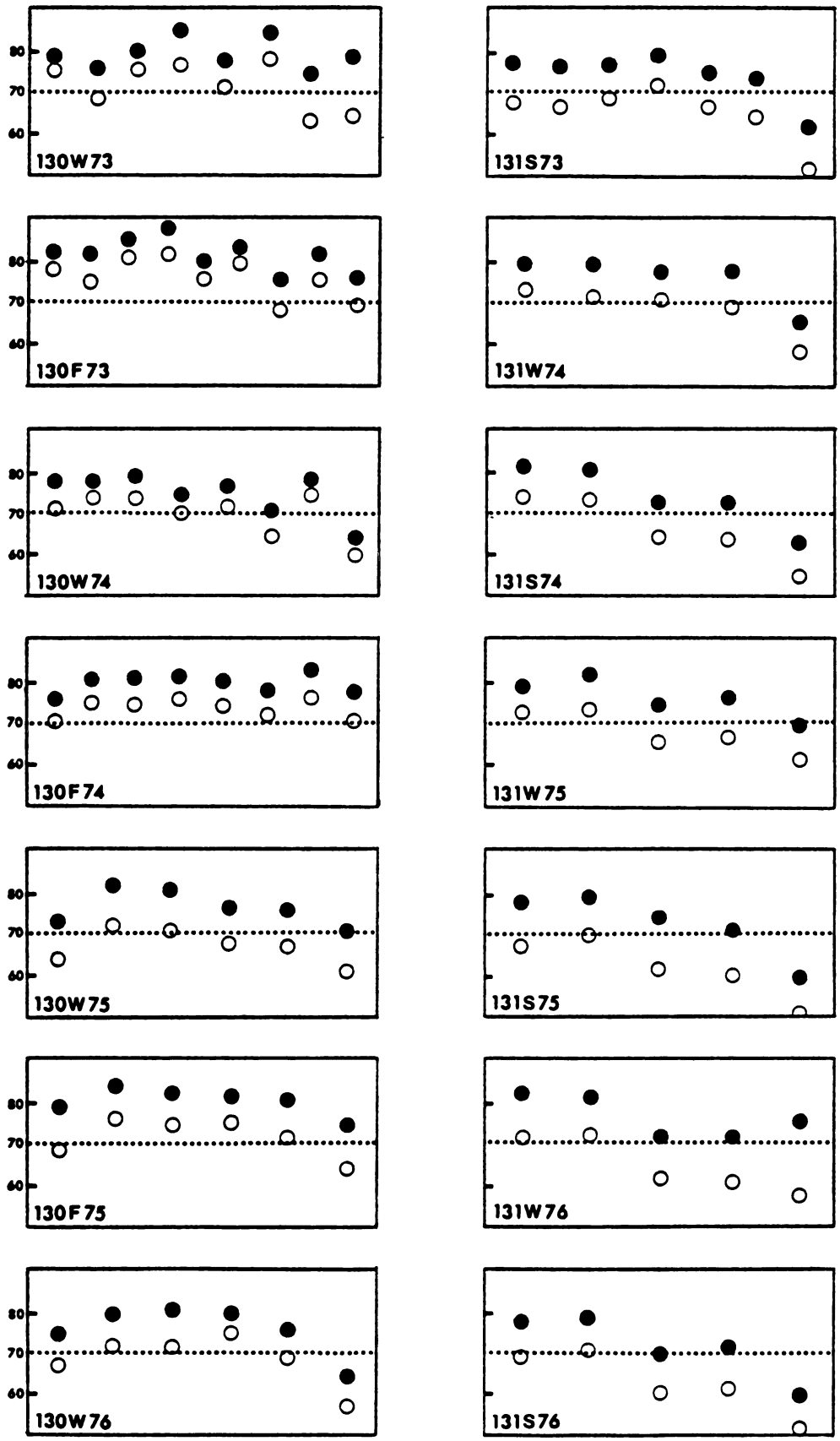


Figure 4.3 Exam mean percents for last tries [●] and all tries [○]

This range of reliabilities has a standard error of from 3.5 to 4.2 percent for a student's term average. Each step in the grading scale spans about seven points, so that only in the case of students near the cutting point for an interval is it likely some grades are incorrectly assigned. Since students are allowed to retake exams, it is probably more likely that any errors in the term average tend toward leniency. However, this leniency in the term average is mitigated by the special effect of the final exam on the term grade.

Representativeness of the tests. The procedures by which tests are constructed is described in Section IV. B. The method of stratified random sampling employed here does not guarantee that all forms of an examination are perfectly parallel but only that what imperfections exist are random and will average out over the course of a term. The effect of these random fluctuations should in almost all cases be near zero in the cumulative term average, thus having virtually no effect on the accuracy of a student's grade.

The grading scale. The selection of where to place the cutting point for each grade interval is influenced by the difficulty of the tests and the expectations of the instructor. The scales used in these courses are based on the assumption that the abilities of students as a group change little from one year to the next, and also that tests used in different terms will be of similar difficulty. Errors in the grading scale affect the entire group and make the course either too hard or too easy. When more students were getting A's and B's than the

instructors considered justified for a typical class*, the scale was adjusted upward by two or three points for each interval. This fine tuning of the scale is necessary because of the impossibility of predicting beforehand the exact difficulties of test questions.

The effect of the final exam. The cumulative final exams given in these courses have a unique effect on the student's grade -- the final examination can only raise or lower the grade one interval. Students are allowed only a single scheduled attempt at the final exam. All students will take the final examination at the same time, and since only one form need be prepared, concerns about the relative difficulty or representativeness of multiple test forms are absent. Because of this, the final exam serves as a correction mechanism within the testing and grading procedure. Students whose term averages are the result of luck or persistence are most likely to have their grade lowered by the final exam. Students whose performance on the cumulative final is sufficiently high have their grade raised because of demonstrated term-end competence. Additionally, the instructor can influence the difficulty level of the final exam and thereby lower or raise the average grade when it appears to be undeservedly high or low. This purposeful manipulation of the test difficulty has not been practiced explicitly in these courses, but the opportunity does exist for its use.

*By Fall 1973 the class average was a full scale step higher than expected -- 3.2 instead of 2.7. When the new scale was implemented in Fall 1974, the average grade dropped to 2.7. This decrease was also due in part to efforts to upgrade the item pool from which test questions are selected.

IV.D. Measuring attitude and attitude change

The method used to measure student attitude at Michigan State University is the survey questionnaire. Multiple-choice questions with two to five response categories are typically used and the student responds on a machine-scorable answer form. As mentioned in the previous chapter, as a result of the unique nature of the teaching methods in CEM 130/131, the standard Student Instructional Rating System (SIRS) forms cannot be used because most of the items on the SIRS form are not relevant. Attitude surveys used in these courses were designed specifically for the modular self-paced method. The two basically different kinds of surveys to which students were asked to respond are the Course Evaluation Form and the Student Precourse (Postcourse) Attitude Form.

IV.D.1. Course evaluation form

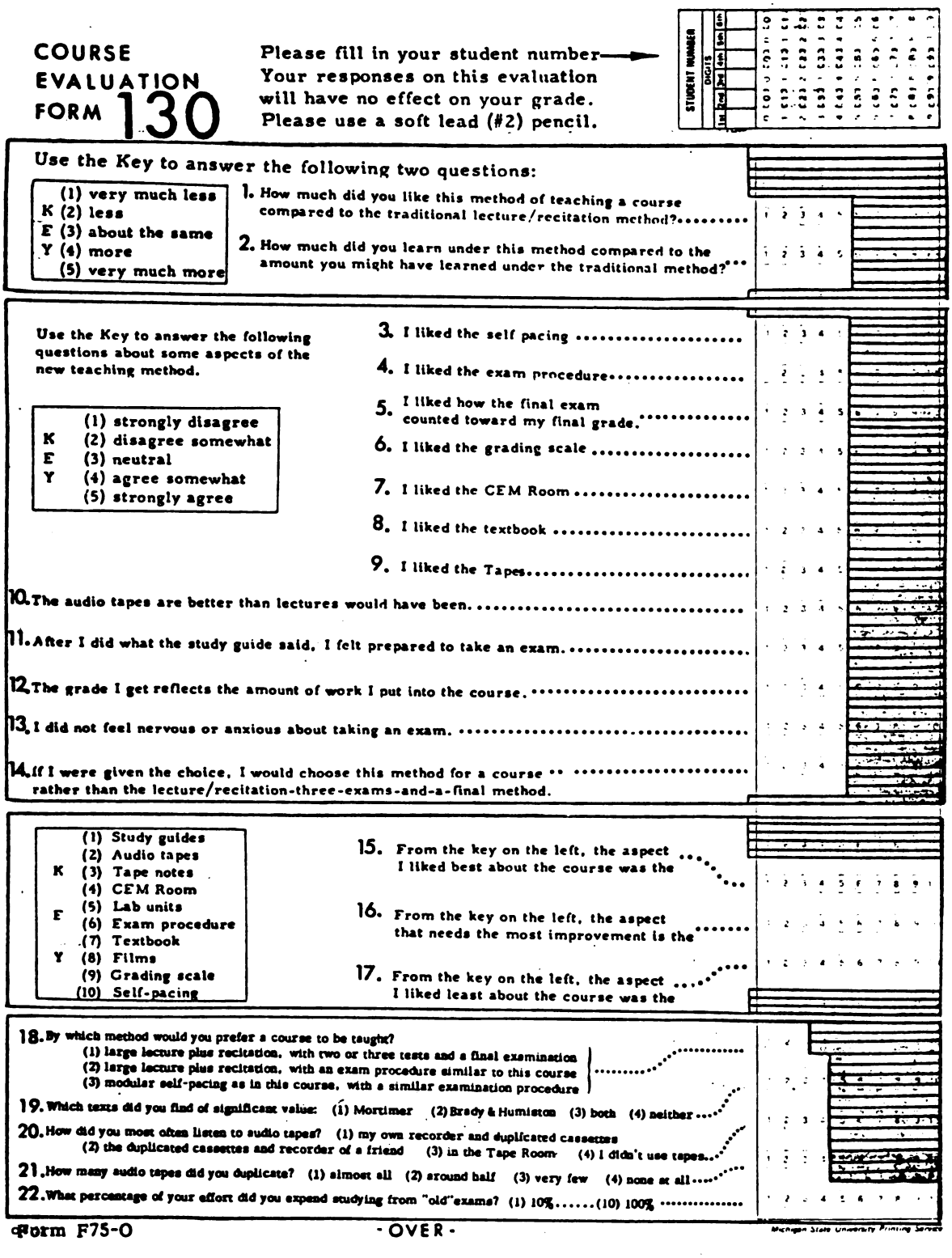
The course evaluation form (Figures 4.4 and 4.5) serves both diagnostic and evaluative functions. Survey items of a primarily evaluative nature are limited to asking the student whether he liked or disliked some particular aspect of the course, or whether some aspect is better than the typical lecture course. In a lecture course it is the person -- not the method -- which the student encounters for the first time. In contrast, most students who take CEM 130/131 have never had a modular self-paced course before; thus the method itself is novel. Because the system is new to the student, all that is surveyed are his

general attitudes toward the elements of the system and not how the implementation of the various elements meets external or accepted criteria. Although it might seem superficial to seek merely a 'happiness quotient', such purpose is sufficient for two reasons.

First, students are not qualified by training or experience to judge many of the aspects of the instructional system. In the case of the lecture-recitation method it may be argued that the weight of the student's experience through twelve years of schooling allows him to evaluate with reasonable accuracy the characteristics of a course and instructor. Since few students have had courses taught by the methods used in these courses, their judgements may be based on unreasonable or atypical expectations instead of being measured against the yardstick of an established norm.

Second, in a very real sense it is irrelevant whether a certain aspect or characteristic of the course is right or good according to some absolute standard. What is relevant is whether the consumer is pleased with the product. Whether the product does what it is meant to do is measured by achievement tests. As Dubin and Taveggio [115] and others [116-118] reported, achievement is not measurably influenced by the method of instruction.* Therefore, the affective goal

*This conclusion is based on the collected literature in the forty years before 1965. The most recent innovations by Keller and others in labor-intensive individualized instruction quite often show statistically significant improvements in student achievement compared with the lecture-recitation method.



COURSE EVALUATION FORM 130

Please fill in your student number → Your responses on this evaluation will have no effect on your grade. Please use a soft lead (#2) pencil.

STUDENT NUMBER — DIGITS: 1st 2nd 3rd 4th 5th 6th

Use the Key to answer the following two questions:

KEY:
(1) very much less
(2) less
(3) about the same
(4) more
(5) very much more

1. How much did you like this method of teaching a course compared to the traditional lecture/recitation method?
2. How much did you learn under this method compared to the amount you might have learned under the traditional method?

Use the Key to answer the following questions about some aspects of the new teaching method.

KEY:
(1) strongly disagree
(2) disagree somewhat
(3) neutral
(4) agree somewhat
(5) strongly agree

3. I liked the self pacing
4. I liked the exam procedure
5. I liked how the final exam counted toward my final grade.
6. I liked the grading scale
7. I liked the CEM Room
8. I liked the textbook
9. I liked the Tapes
10. The audio tapes are better than lectures would have been.
11. After I did what the study guide said, I felt prepared to take an exam.
12. The grade I get reflects the amount of work I put into the course.
13. I did not feel nervous or anxious about taking an exam.
14. If I were given the choice, I would choose this method for a course rather than the lecture/recitation-three-exams-and-a-final method.

KEY:
(1) Study guides
(2) Audio tapes
(3) Tape notes
(4) CEM Room
(5) Lab units
(6) Exam procedure
(7) Textbook
(8) Films
(9) Grading scale
(10) Self-pacing

15. From the key on the left, the aspect I liked best about the course was the
16. From the key on the left, the aspect that needs the most improvement is the
17. From the key on the left, the aspect I liked least about the course was the

18. By which method would you prefer a course to be taught?
(1) large lecture plus recitation, with two or three tests and a final examination
(2) large lecture plus recitation, with an exam procedure similar to this course
(3) modular self-pacing as in this course, with a similar examination procedure
19. Which texts did you find of significant value: (1) Mortimer (2) Brady & Humiston (3) both (4) neither
20. How did you most often listen to audio tapes? (1) my own recorder and duplicated cassettes (2) the duplicated cassettes and recorder of a friend (3) in the Tape Room (4) I didn't use tapes.
21. How many audio tapes did you duplicate? (1) almost all (2) around half (3) very few (4) none at all
22. What percentage of your effort did you expend studying from "old" exams? (1) 10% ..... (10) 100%

Form F75-O - OVER -

Figure 4.4 Sample course evaluation form [front] from CEM130F75

Please make any comments you have about the course. How do you feel now about this method compared to how you felt at the beginning of the term? Are there ways in which this course could be improved? What aspects are there which you would like to see continued in this course and expanded to other courses?

Please make detailed comment about the Audio Tapes - their usefulness, quality, pacing, use of examples, etc.:

Figure 4.5 Sample course evaluation form [back] from CEM130F75

being tested by many items on the survey is whether the student liked the method and thought it as good as or better than the lecture-recitation method.

The last section of the course evaluation form contains several diagnostic or informative survey items. These multiple-choice items have varied for different terms, as 'How often have you used an electronic calculator?' or 'How many audio tapes did you duplicate?' Such items poll student behavior and attitude on specific issues, and the results have influenced actual or projected course operation.

IV.D.2. Precourse and postcourse surveys

The method chosen to measure attitude change is the before-and-after survey of student opinion on precourse and postcourse attitude forms. These forms (Figures 4.6 and 4.7) are different only when the tense of a sentence has to be either future expectation or past experience. Three variables are compared across the time interval of the courses: (1) whether chemistry is enjoyable and a chemistry course interesting, (2) whether the course is useful, and (3) how different learning situations compare in effectiveness. Many other questions of ambiguous classification were included on the forms when they were first designed. These and other items remain on the forms so that successive terms may still be compared under the same survey conditions, although the results of the items are not reported.

The general format for survey items on the precourse and postcourse attitude forms is descriptive multiple-choice (cf. Ch. 3). This type of item was chosen because of its flexibility; the content or wording of a question need not be unduly manipulated to fit an invariant answer key. The principal drawback to the descriptive answer key is that the numbers assigned to the response categories may not actually possess the numerical properties implied. In order that calculation of means and deviations may be simplified, it is assumed that the response categories are interval and not merely ordinal or nominal. (This routinely made assumption in attitude survey research is so seldom strictly true that special caveat should be made more often.)

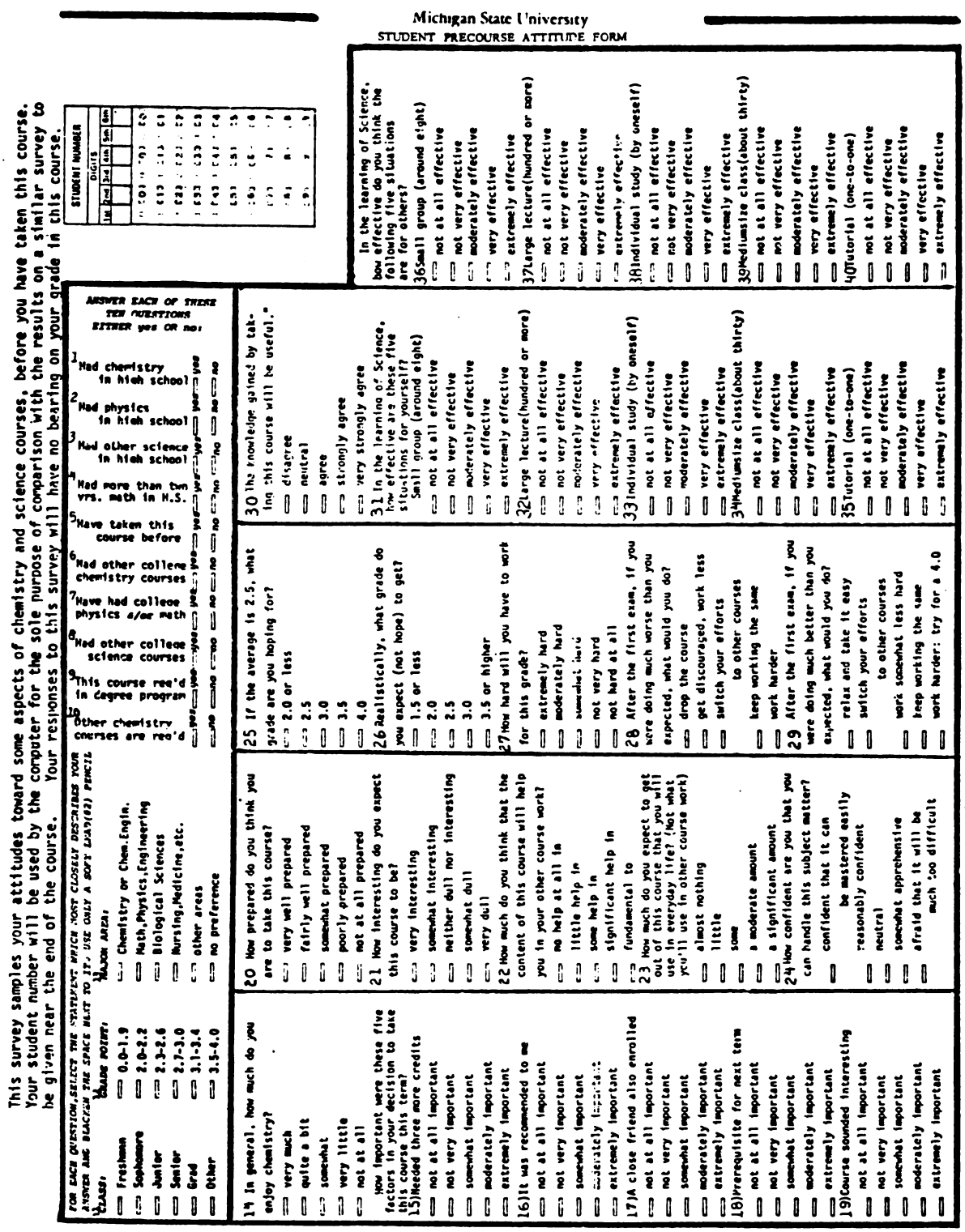


Figure 4.6 Sample precourse attitude form

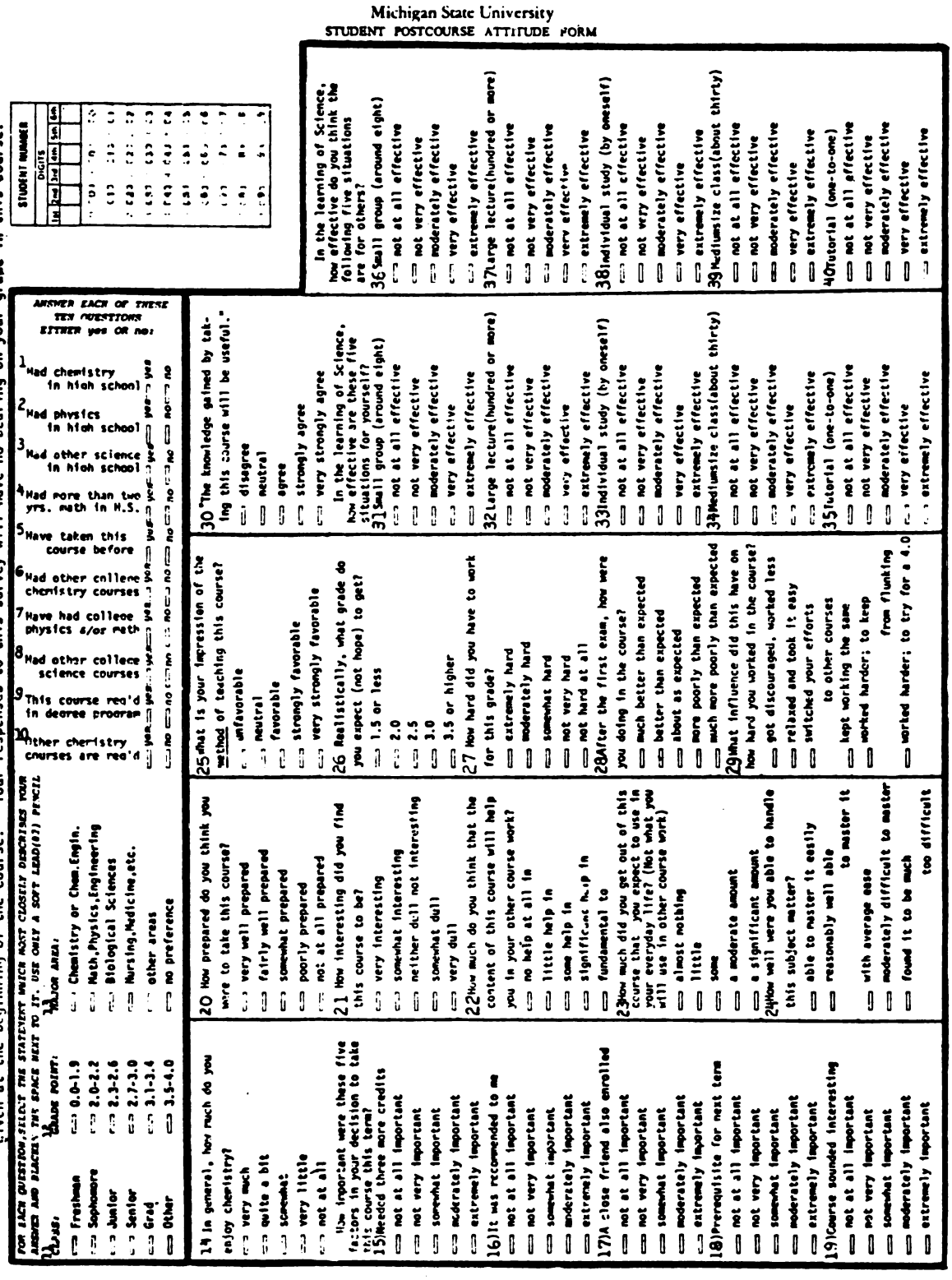


Figure 4.7 Sample postcourse attitude form

IV. E. What are the student attitudes?

Attitude surveys, like achievement tests, are a collection of questions given to a collection of students. Four different kinds of numbers can be reported: a single student's response on a single item, the group mean for an item, a student's total score on all items, and the group mean across all items. The response of a single student to a single item is seldom of surpassing interest. The group mean for an item is the response mean on an attitude survey or the item difficulty on an achievement test; both are of interest.

The total score on all items for a single student is meaningful only when the entire test or survey is of sufficiently homogeneous content that subtest scores are unnecessary. Rarely are the questions on an attitude survey similar enough to justify the calculation of a total score. Even many achievement tests are divided into subtests such as mathematics and verbal, or numerical, natural science, social science, verbal, and general. The subtests on an attitude survey are more often discussed as factor scores. Through statistical factor analysis, certain survey questions are grouped together and a mean response for a factor may be reported. Because more than one item is included in a factor score, the mean response is more stable than that of a single question. Still, the attitudes of a single student may vary more from day to day than his level of achievement and therefore the responses of individual students are of less value than the corresponding group mean on an item or factor.

The mean on an achievement test is the score for the entire group on all questions. Since attitude surveys are typically more heterogeneous in content, the mean factor (subtest) score is a more desireable statistic. Therefore in this study only the factor means and item mean scores of the course evaluation and of the precourse and postcourse attitude surveys are presented. Inasmuch as these data are reported at the group level, and because one hundred percent of the students do not respond to each attitude survey or evaluation, the response rate must be considered and its possible influence on the observed means examined.

Favorability of response and rate of return. Whenever the response rate is not one hundred percent, the results from the actual sample of students who respond may differ from what the results might have been had all students responded. Whether the inclusion of the nonresponding students would raise or lower the observed mean depends on whether the responses of these students would be more or less favorable than those of the sample which did respond.

Several reasons for nonresponse are possible. Some students may withhold the cooperation requested on an evaluation form because of their dislike of the course. Other students may feel favorably disposed, but be too apathetic to respond. Still others may spend every minute of the final exam period on the test and have no time remaining in which to respond to the surveys. In most cases when the response is very nearly complete, these different types of nonresponse

have a negligible cumulative effect on the group mean. Yet this assumption cannot be directly verified since the responses of the students who do not respond are unknown.

Correcting for incomplete survey return. An indirect test of the effects of nonresponse was made by calculating the correlation between rate of return and response mean. If there is no relationship between rates of return (which vary between fifty-five and ninety percent), it will be assumed that the observed group mean is no different from the true total group mean. Factor 11 from the course evaluation form was selected for this test because it is a strong factor composed of five items which indicates the students' general attitude toward CEM 130/131. The correlation coefficient between percent return and favorability of response was found to be 0.69. This unexpectedly high correlation indicates that students who do not respond probably would have responded favorably on the surveys. The regression line of mean response on percent return and the observed and corrected response means for Factor 11 are presented in Figures 4.8 and 4.9 on the next page.

Obviously the percent survey return is not the cause of student attitudes. Yet it is also clear from these data that incomplete survey response distorts the mean in a negative direction; the greater the short-fall from one hundred percent, the less favorable the observed response. In light of this, all observed group means should be considered lower bounds to the favorability index. Regressions and

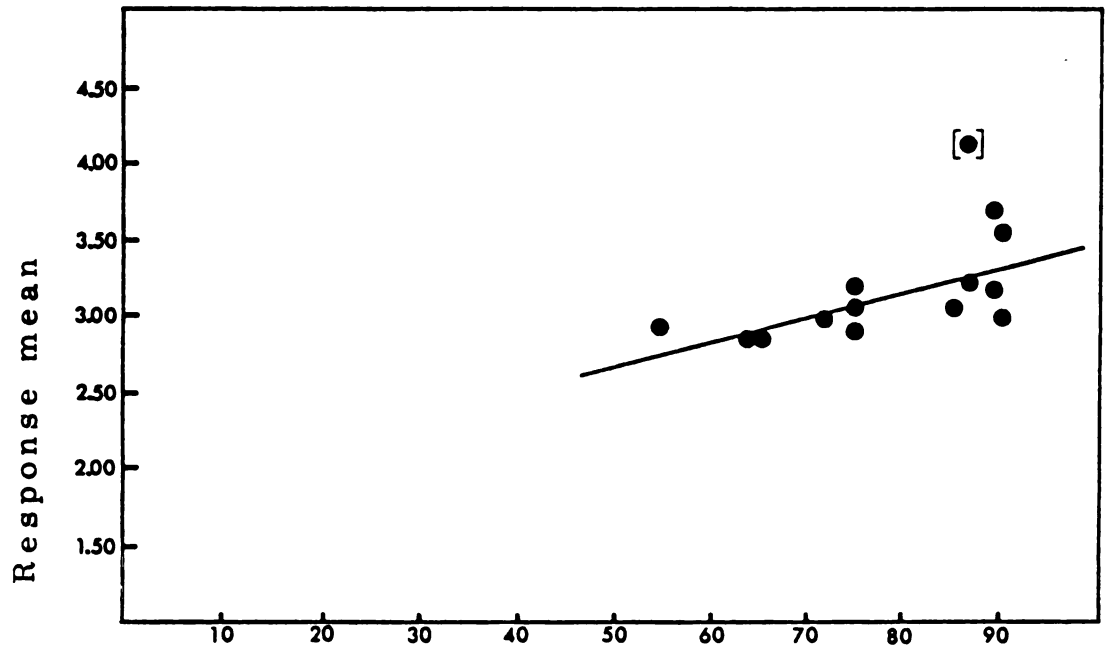


Figure 4.8 Regression plot of mean response on percent return; mean response = 1.877 + 0.01606 (percent return)

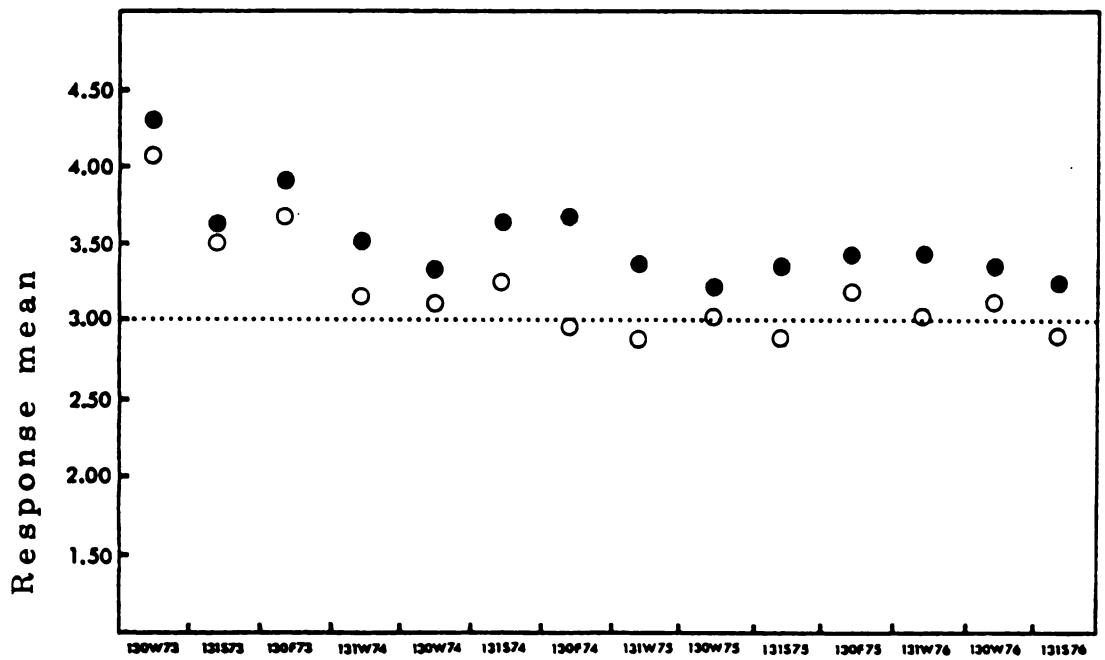


Figure 4.9 Observed [○] and adjusted [●] response means for Factor 11 from the course evaluation form

adjustments are not calculated for other items and factors because the equations are imperfect, and the true correlation for any single group may be different from the average correlation suggested by the equations. It was deemed appropriate not to add any further imprecision or inaccuracy to these data by adjusting them with an imperfect transformation. Nevertheless it must be emphasized that the reported means are lower bounds to the true group attitudes.

IV.E.1. Results of the course evaluation

A sample course evaluation form distributed at the end of each term is displayed in Figures 4.4 and 4.5. Graphs displaying the means, medians, and approximate fractions of the group choosing each resp response category are compiled in Appendix F. A graph of mean responses for the two strongest factors generated by factor analysis is presented in Figure 4.10. Factor 11 is a general 'happiness quotient' composed of Items 1, 2, 3, 10, and 14 from the evaluation form. Factor 12 denotes specific satisfaction with the examination procedure and includes Items 4, 5, and 6 from the survey. The major choices among ten course aspects in the forced-choice poll requested by Items 15, 16, and 17 are listed in Table 4.1.

The overall attitude of the students toward the course is neutral, but this is an average of some aspects that students like very much and others some students dislike intensely. Students consistently like best the testing procedures used in CEM 130/131, followed by the self-pacing.

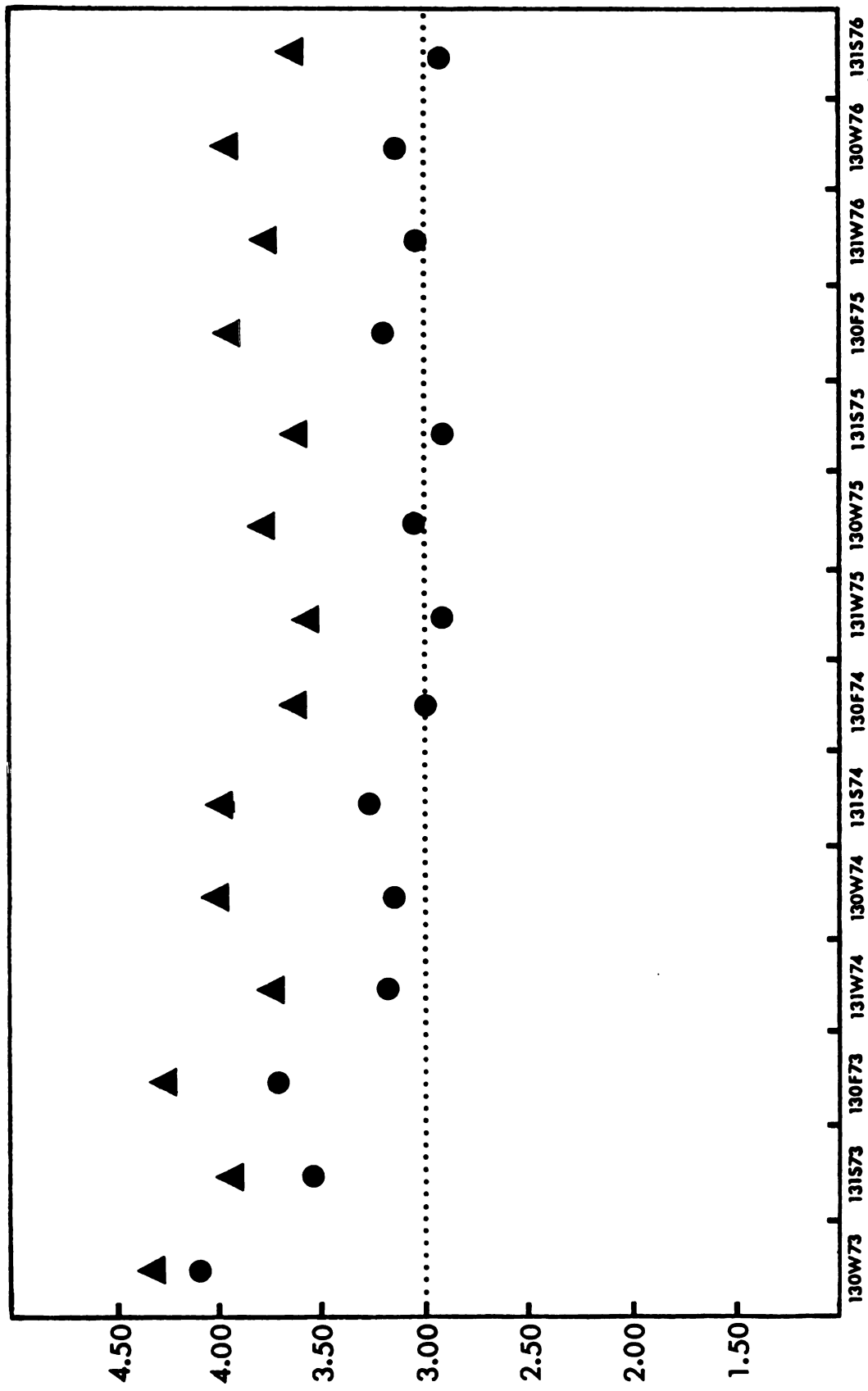


Figure 4.10 Response means for Factor 11 [●] and Factor 12 [▲] from course evaluation form

Table 4.1 Course evaluation response percents to Items 15, 16, and 17

Term	Item	Study Guides	Audio Tapes	Tape Notes	CEM Room	Lab Units	Exam Proc.	Text	Films	Grade Scale	Self-pacing
	15		*12*				*35*				*31*
130W73	16				30	19		14			
	17				10	28		19	12		9
	15						*37*				*23*
131S73	16	15			22			16			
	17				14			20			8
	15										
130F73	16										
	17										
	15						*34*			*13*	*26*
131W74	16		13		46						
	17		13		29			13			12
	15						*46*				*24*
130W74	16		18		40						
	17		18		23			11	14		6
	15						*42*				*22*
131S74	16	11	16		26			13			
	17		15		14	13		15			11
	15						*35*				*29*
130F74	16		22		19			15			
	17		17		10			16	10		15
	15						*32*				*30*
131W75	16		16		32			11			
	17		18		19			16			13
	15						*42*				*25*
130W75	16		25		27			12			
	17		20		17			17			11
	15						*40*				*23*
131S75	16		19	10	26			14			
	17		23		14			18			7
	15						*42*				*29*
130F75	16		23	11	14			17	10		
	17		23					19	16		10
	15						*41*				*23*
131W76	16	10	20		28			11			
	17		22		15			15			6
	15						*45*				*30*
130W76	16	12	18	12	25			14			
	17		20		16			18	10		7
	15						*47*				*19*
131S76	16	11	25	12	20			11			
	17		26	11	15			13			9

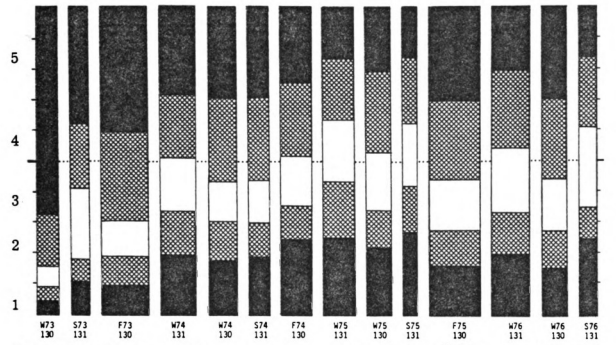


Figure 4.11 Item 14 from the course evaluation: "If I were given the choice, I would choose this method for a course rather than the lecture/recitation-three-exams-and-a-final method."

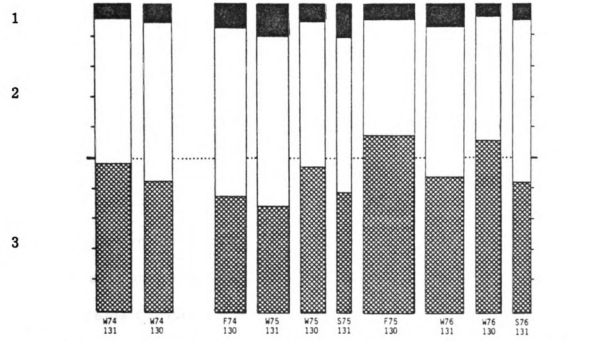


Figure 4.12 Item 18 from the course evaluation: "By which method would you prefer a course to be taught?"

Dislike is more equally divided between audio tapes, CEM Room, textbook, and self-pacing. Thus the modular self-pacing feature of these courses is, for different students, both the most and least liked aspect. This polarization of attitude means that the more the instructor pleases one group, the more he will displease the other group.

Given the choice of instructional method, the balance of student attitude tips in favor of modular self-pacing by which these courses are taught as opposed to a traditional lecture method. Yet the true second choice of the group is not the typical lecture course but a hybrid combining the live lecture and the self-paced examination procedure. The distribution of student responses among these choices is illustrated in Figures 4.11 and 4.12.

It is understandable that given a choice between the rigidly structured lecture course and this more flexible design that less than forty percent of the students clearly prefer the lecture method. What is surprising (and gratifying) is that when students are given the option of a method which includes the most popular aspect of this system -- the exam procedure -- and eliminates the least popular aspect by substituting live lectures for audio tapes, that nearly half of the students still prefer the modular self-paced method 'as is'. The very small fraction of students who select a straight lecture method gives some indication of the underlying unpopularity of large lectures.

Instructor interaction with students and evidence of his interest in their welfare has been reported [119-121] many times as one of the most important and desirable characteristics of a course. Printed study guides, prerecorded audio tapes, and computer generated exams all lack a live person with whom to interact. It is also difficult to project through these mechanisms the interest an instructor has in his students' welfare. Written and verbal comments made by students indicate that the impersonality of these course aspects is what displeases them most. So too does the impersonality of the large lecture course dull its competitive appeal for students -- if a student is going to be merely a face in the crowd, he will at least be able to choose his own study and testing times in this method. Perhaps the abrupt change from a high school classroom in which an instructor may know each of his students almost personally to the mass-education setting of a university engenders a sometimes violently negative attitude toward the whole system.

IV.E.2. Patterns in student attitude

Three relationships among student attitudes and with other variables are of interest. First, the intercorrelations of attitude survey questions leads to elucidation of underlying general attitude factors. Second, responses to surveys at both the beginning and end of a course indicate changes in student attitude. Third, how well a student performs in a course influences his attitudes toward and evaluation of

the course and the subject. The results of the investigation of each of these relationships will now be discussed.

IV. E. 2. a. General attitude factors

A single question on an attitude survey allows a student to respond with perhaps only five choices. A chemistry test with four equivalent questions provides as many possible scores as a survey item with five response categories. Just as the restricted range of such a chemistry test gives an unreliable estimate of achievement, so too does the single survey item provide a less than perfectly reliable estimate of attitude. Yet fewer attitude questions are required to obtain a reasonably accurate score for a student than would be required for an achievement test. Three or four equivalent survey questions provide a response range about equal to that of a fifteen-question examination. The two surveys used in these courses contain twenty-two and forty questions. It might be expected that more than one item on a survey actually measures a single attitude by asking the same question in different words. The procedure used to identify these equivalent items (and thereby to identify the underlying attitude measured by them) is factor analysis.

Factor analysis of the attitude surveys. Factor analysis is a post hoc method of determining the correlations among test or survey questions and of grouping them into sets of equivalent items called factors. The wording and content of all items in a factor are the basis

for recognizing the underlying attitude which each of the questions measures. The arcane details of the correlation matrices and item loadings are not presented here. Those factors with the greatest precision and strongest identities are listed in Table 4.2. The mean responses to each factor are presented in Appendix F, and in Table 4.6 on p. 240.

Factors such as 2 or 12 are composed of questions which are clearly similar in content; such easily discerned relationships between items will not be discussed. The most provocative and enlightening factors contain survey questions about the effectiveness of different learning situations. The student is asked to rate the effectiveness of five learning conditions: individual study, tutorial, small groups of around eight, medium sized classes of about thirty, and large lectures of a hundred or more. Students are also asked to differentiate between personal effectiveness and effectiveness for others. It was expected that these five choices are different enough to elicit separate responses from students. It might also be expected that students differentiate between situations which are personally effective and situations which are generally effective for most other students. Both expectations were unrealized.

Students did not distinguish between conditions which they find personally effective and those which most other students find effective. This may indicate either that students are not as ideosyncratic in their learning patterns as might be thought, or that they are unable to judge

Table 4.2 Major factors from the attitude surveys

Factor*	Description	Survey questions
1 & 6	Enjoyment and interest	14. In general, how much do you enjoy chemistry? - very much - quite a bit - somewhat - very little - not at all 19. Course sounded interesting - not at all important - not very important - somewhat important - moderately important - extremely important 21. How interesting do you expect this course to be? - very interesting - somewhat interesting - neither dull nor interesting - somewhat dull - very dull
11	General favorable attitude toward the course	1. How much did you like this method of teaching a course compared to the traditional lecture-recitation method? - very much less - less - about the same - more - very much more 2. How much did you learn under this method compared to the amount you might have learned under the traditional method? - (same response categories as 1) 3. I liked the self-pacing. - strongly disagree - disagree somewhat - neutral - agree somewhat - strongly agree 10. The audio tapes are better than lectures would have been. - (same response categories as 3) 14. If I were given the choice, I would choose this method for a course rather than the lecture-recitation-three-exams-and-a-final method. - (same response categories as 3)
12	Liked testing procedures	4. I liked the exam procedure. - (same response categories as 3) 5. I liked how the final exam counted toward my final grade. - (same response categories as 3) 6. I liked the grading scale. - (same response categories as 3)

*Factors 1 through 10 encompass items from the precourse and postcourse attitude surveys; factors 11 and 12 pertain to the course evaluation form. The factors from the precourse and postcourse surveys are identical in content; factors 1 and 6, 2 and 7, and so forth are pre- and post- equivalent pairs.

Table 4.2 (cont'd)

Factor	Description	Survey questions
2 & 7	Usefulness	23. How much do you expect to get out of this course that you will use in everyday life? (Not what you'll use in other course work) - almost nothing - little - some - a moderate amount - a significant amount 22. How much do you think that the content of this course will help you in your other course work? - no help at all in - little help in - some help in - significant help in - fundamental to 30. "The knowledge gained by taking this course will be useful." - disagree - neutral - agree - strongly agree - very strongly agree
3 & 8	Effectiveness of very small group learning situations	In the learning of Science, how effective are these ... situations ... for yourself? 31. Small group (around eight) - not at all effective - not very effective - moderately effective - very effective - extremely effective 35. Tutorial (one-to-one) - (same response categories as 31) ... situations ... for others? 36. Small group (around eight) - (same response categories as 31) 40. Tutorial (one-to-one) - (same response categories as 31)
4 & 9	Effectiveness of large group situations	... situations ... for yourself? 32. Large lecture (hundred or more) - (same response categories as 31) 34. Medium size class (about thirty) - (same response categories as 31) ... situations ... for others? 37. Large lecture (hundred or more) - (same response categories as 31) 39. Medium size class (about thirty) - (same response categories as 31)
5 & 10	Effectiveness of self-study	... situations ... for yourself? 33. Individual study (by oneself) - (same response categories as 31) ... situations ... for others? 38. Individual study (by oneself) - (same response categories as 31)

the general effectiveness of a situation separate from its effectiveness for themselves. Thus in every case, the parallel questions about a particular learning situation '...for yourself' and '...for others' are both in the same factor.

Students do not consider the five learning conditions to be distinctly different. The clustering of survey questions by their intercorrelations leads to only three categories: self-study, small group, and large group situations. It appears that the definition of a large group has a lower limit of as few as thirty; the lower bound may be even smaller. It also appears that one-to-one tutorial instruction is judged to be about as effective as a group of eight students with an instructor. Again, eight students is a limit (in this case an upper limit), and the size of group instruction equal in effectiveness to tutorial instruction may in fact be larger.

This extreme dichotomization of class sizes into small and large, with the transition somewhere between eight and thirty, has practical application. If the instructor desires to provide the kind of learning environment described as tutorial, he need not actually establish a one-to-one student-teacher ratio -- a ratio as high as eight-to-one is approximately as effective from the student's point of view. Conversely, unless the class size can be brought considerably under thirty, it might just as well be a large lecture. Perhaps students perceive no middle ground between the individualized instruction available in very small groups and the impersonality of larger lecture classes. In actuality

there may be some intermediate class size which has characteristics distinct from tutorial or large lecture classes, but these data indicate only that if such a size exists it must lie between eight and thirty students. However, there is no direct evidence in these results which portend such a midsize class.

IV.E.2.b. Precourse and postcourse response differences

The student attitudes measured by the attitude survey given at the beginning of a course may differ from those measured at the end of a course. To investigate such differences, student attitudes were surveyed at the beginning of CEM 130, at the end of CEM 130, and at the end of CEM 131. There are only five samples with percent survey returns high enough and close enough together to give any indication of pre-post effects not unduly clouded by response rate effects. These five samples are displayed in Table 4.3.

Some tentative generalizations about and possible explanations of these results may be made. The overall trend seems to be a decrease in the favorability of student attitudes on all factors with the exception of Factor 10. Almost all students who respond to these surveys are freshmen newly enrolled in the university. CEM 130 is the first chemistry course these students are taking, and they might have bright visions of the enjoyability, usefulness, and effectiveness of the course which do not stand the test of time.

Table 4.3 Differences in mean response on postcourse attitude factors compared with precourse responses

Factor	130W73	131S73	130F73	130W75	130F75	mean
6	-.01	-.46	-.12	-.05	-.26	-.11
7	-.49	-.88	-.48	-.60	-.59	-.54
8	-.12	+.02	-.10	-.02	-.24	-.12
9	-.20	-.14	-.12	-.04	-.06	-.10
10	+.15	-.04	+.02	-.03	-.04	+.03
precourse return	80		88	68	78	
postcourse return	71	73	78	65	80	

Factor 7 pertains to the usefulness of chemistry and always decreases dramatically from the precourse to the postcourse survey. Over ninety percent of the students in these courses are nonmajors and may view much of what they are required to learn as unrelated to their major but of use only to pass chemistry tests. Students may not recognize the utility of the subject matter until later in their academic careers. Perhaps in future multipath versions of these courses more closely allied to non-chemistry majors, students will view the subject as more useful.

Factors 8, 9, and 10 are ratings of the effectiveness of small classes, large lectures, and individual study. The student ratings of individual study do not show the decrease in response mean exhibited

for ratings of small classes and large lectures. Since these surveys are given before and after an individual-study method, it may be concluded that the student's immediate experience with such a method has convinced him of its effectiveness. A second more tentative and indirect conclusion may be drawn. The postcourse responses to Factors 8 and 9 may be lower than the precourse means not because students feel that such methods are less effective than previously thought but because the less than complete rates of return and generally lower returns on the postcourse surveys have weighted the unfavorable viewpoint. Based on this assumption that responses on Factors 8 and 9 should have been the same, corrected approximate changes in attitude may be estimated; these are displayed in Table 4.4

Table 4.4 Corrected approximate changes in student attitude

Factor	Description	Change in response
6	enjoyment and interest	~ 0.00
7	usefulness	~ -0.43
8	effectiveness of small groups	~ 0.00
9	effectiveness of large lectures	~ 0.00
10	effectiveness of individual study	~ +0.13

The above approximations, though of dubious pedigree, support the previously offered interpretations of the attitude survey results. Students feel that the material covered in these courses is considerably

less useful than they expected when they began. Students also view the effectiveness of individual study at least as favorably as they did before taking CEM 130/131.

IV.E.2.c. Relation between attitudes and grade

Do students manifest positive attitudes toward a course when they perform well in it and negative attitudes when they perform poorly? This question has been the topic of much discussion and research. Unfortunately, the answer is equivocal. Numerous studies [121-130] of sometimes subtly differing methodology have shown correlations between grades and attitudes ranging from extremely negative -- poor performance and favorable evaluation -- to moderately positive. The consensus of results is that the relationship is weakly positive and often nonsignificant. For example, Remmers [125] found a correlation of only 0.24 between chemistry grades and student ratings of their chemistry courses. Cornwell [124] found correlations ranging from 0.20 to 0.37 typical of university courses.

Bausell [129] investigated not only the relationship of attitude with grade but also with expected grade and overall college grade point average. Embedded in the usual near-zero correlation of grade with attitude, he found that when the grades students received were much different from their all-university averages, the relationship was positive -- i.e., if a student received a higher grade than his average course grade, he rated that course more favorably, and if he received

a lower-than-average grade he rated the course less favorably. Thus the attitudes of the group or students must be considered not only in relation to achievement in the course, but also in relation to the typical performance in other college courses. If the grades in a course are generally lower than grades in other courses, the favorability of ratings will also be lower.

The response means for factors from the attitude and evaluation surveys and their correlations with course grade ate presented in Table 4.6. The correlations between precourse attitudes and final grades are uniformly near zero.* In contrast, all but two of the other factors show generally positive correlations with grades. Correlations of postcourse and evaluation factors with grades are listed below.

Table 4.5 Mean correlation of attitude factors with course grade

Factor	Description	$\bar{r}$	σ_r	n
1	enjoyment and interest	.30	.04	12
2	usefulness	.28	-.05	12
3	effectiveness: small groups	.03	.04	12
4	effectiveness: large groups	-.08	.05	12
5	effectiveness: self study	.24	.05	12
11	general favorable attitude	.30	.07	12
12	liked testing procedures	.22	.08	12

*$\bar{r}$ = .03, σ_r = .05, n = 35

Table 4.6 Attitude factor means and correlations with course grade

Course	One*		Two		Three		Four		Five		Eleven	Twelve
	pre	post	pre	post	pre	post	pre	post	pre	post	eval	eval
130W73	3.27	3.27	3.64	3.15	4.30	4.18	3.05	2.84	3.29	3.44	3.18	3.89
	.08	.29	.02	.20	-.03	.12	.07	-.09	.00	.20	.24	.22
131S73		2.81		2.76		4.31		2.90		3.25	3.16	3.96
		.33		.28		.10		-.13		.21	.41	.20
130F73	3.42	3.30	3.67	3.19	4.39	4.30	3.10	2.97	3.32	3.34		
	.02	.21	.03	.25	.00	.05	.05	-.03	-.02	.16		
131W74		2.85		2.78		4.30		3.10		3.17	3.07	3.82
		.34		.32		-.03		-.05		.24	.26	.19
130W74	(3.02)		(3.53)		(4.47)		(3.05)		(3.39)		3.09	4.01
	(.00)		(-.04)		(.19)		(.05)		(.18)		.18	.14
131S74		2.78		2.72		4.22		2.96		3.26	3.09	3.96
		.36		.37		.02		-.02		.29	.33	.17
130F74	3.35	(2.95)	3.71	(2.86)	4.45	(4.30)	3.18	(3.16)	3.28	(3.15)	2.99	3.63
	.10	(.32)	.02	(.33)	.00	(-.03)	.05	(-.06)	.00	(.26)	.19	.11
131W75		2.81		2.75		4.30		3.15		3.12	3.00	3.55
		.26		.28		.02		-.06		.21	.29	.14
130W75	3.02	2.96	3.43	2.83	4.33	4.31	3.11	3.07	3.21	3.18	3.12	3.86
	.05	.29	-.01	.30	.02	.02	-.01	-.15	.01	.24	.32	.22
131S75		2.66		2.56		4.35		3.07		3.17	3.02	3.65
		.26		.20		.05		-.17		.36	.37	.25
130F75	3.34	3.08	3.70	3.11	4.45	4.30	3.17	3.09	3.28	3.24	3.19	3.96
	.04	.28	.02	.24	.06	.04	-.03	-.07	.06	.23	.34	.29
131W76		2.70		2.72		4.35		3.15		3.19	3.08	3.76
		.30		.32		.00		-.11		.22	.39	.33
130W76	3.06	2.94	3.41	2.90	4.39	4.34	3.12	3.01	3.21	3.26	3.17	3.95
	-.02	.29	-.04	.31	.03	-.02	.07	-.06	.02	.21	.36	.38
131S76		2.67		2.58		4.32		3.05		3.14	3.05	3.69
		.34		.31		.04		-.04		.25	.28	.18

*For each term the mean response on a scale of 1 to 5 for each factor is given in the first row, and the correlation of the factor with the course grade is given in the second rwo. The compositions of the factors are:

One	Enjoyment and interest	Pre-post Nos. 14, 19, 21
Two	Usefulness	Pre-post Nos. 22, 23, 30
Three	Effectiveness of very small group learning situations	Pre-post Nos. 31, 35, 36, 40
Four	Effectiveness of large group situations	Pre-post Nos. 32, 34, 37, 39
Five	Effectiveness of self-study	Pre-post Nos. 33, 38
Eleven	General favorable attitude	Evaluation Nos. 1, 2, 3, 10, 14
Twelve	Liked testing procedures	Evaluation Nos. 4, 5, 6

Factors 3 and 4 pertain to different methods of instruction and it is not surprising that they are uncorrelated with how well a student performs in these courses. The other factors all show a positive relationship between favorability of response and course grade indicating that students who did well in the course responded more favorably on theattitude and evaluation surveys than did students who did poorly.

The grades assigned in CEM 130/131 tend to be lower than the overall university course grade.* In light of Bausell's work, it would then be predicted that the student responses on attitude questionnaires would show a positive correlation with course grade, and that the mean response would be lower than typical for university courses. Both of these effects are observed.

Grading trends in CEM 130/131. The mean grades received in CEM 130/131 are displayed in Figure 4.13. Squares identify CEM 130; circles CEM 131; triangles are the all-university freshman GPA. Open symbols designate the mean grade received in all courses taken by students enrolled in CEM 130 or 131 and are provided for selected available terms; closed symbols represent the mean course grade in CEM 130 or 131. Course enrollments of fewer than 300 students are

*The grades received in CEM 130/131 average about 0.24 unit lower than the mean grade received in all courses taken by students in CEM 130/131 that term. The average chemistry grade was higher than the term average only six times in thirty available term comparisons; all these positive differences occurred in CEM 130. Negative differences occurred most frequently and with greatest magnitude in CEM 131.

classed as small [■,●], as medium [■,●] with between 300 and 850 students, and as large [■,●] with more than 850 students.

When CEM 130 and 131 were taught with the traditional method, the conversion of test scores into grades was done according to the expectations of the instructor and his inspection of a scatter sheet. When the difficulty and equivalence of examinations varied, the grading scale was corrected for such variation by a post hoc grading curve. This method may impose a uniformity in grade means which is unjustified or may exaggerate differences between terms when instructors with different perceptions of group abilities and the distribution of test scores prepare the grading scale. In contrast, the trends in grades under the preset grading scales used in the new method are more the result of differences in students rather than of differences between instructors.

Two trends worthy of mention appear in these term averages between the traditional and the modular methods. The first trend is the general decrease in grades for summer session courses. The second trend is an increase in mean grade in CEM 130 contrasted with a decrease in CEM 131 so that now the CEM 130 means are generally higher than the CEM 131 grade means.

Summer term grades in the new method are based on grading scales and examinations directly equivalent to those used during other terms. Different group abilities and not the different teaching and testing strategies of a summer instructor are the cause of the observed

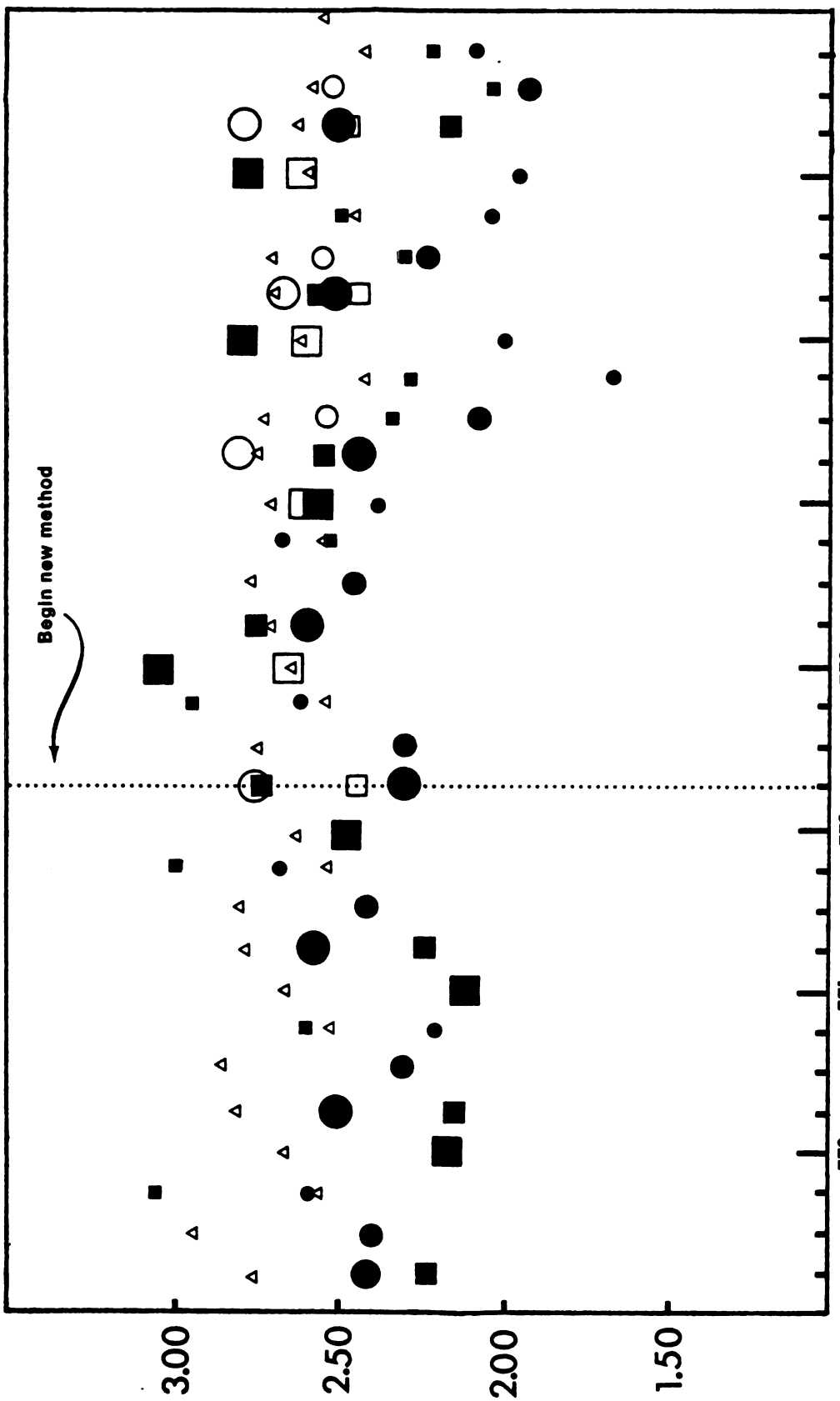


Figure 4.13 Mean course grade in CEM 130 [■] and CEM 131 [●] and freshman GPA [▲]; see text.

grade distributions. No allowances are made in test or grading scale difficulties for those students (about half the summer enrollments) who take both courses in one term. Because more tests are given in the new method than before, students taking both courses must study for eleven separate exams within the ten week period.

Average grades in CEM 131 are very slightly lower now than in the lecture course, while grades in CEM 130 are significantly higher than before. These enrollment-weighted mean grades are listed in Table 4.7.

Table 4.7 Average course grades weighted for enrollment

Course	Old mean	New mean
130	2.308	2.668
131	2.431	2.363

The current order of the grade means -- CEM 131 lower than CEM 130 -- is more realistic since CEM 131 is believed to be a more difficult course than CEM 130.

IV. F. Summary of achievement and attitude

Student achievement in CEM 130/131 follows a remarkably consistent annual pattern ranging from 78 percent for fall terms of CEM 130 to 69 percent for spring terms of CEM 131. The total test means, last-try means, and mean course grades are always lower in CEM 131 than in CEM 130. This decrease may be because the material in CEM 131 is more sequential, and success in later units requires a grasp of earlier concepts. In contrast, CEM 130 contains much descriptive material, and success in later units often depends only on the memorization of new facts somewhat unrelated to previous material.

The average grade assigned in CEM 130 is 2.80 and in CEM 131 it is 2.46. The only term in which the mean grade does not parallel the mean test scores is Fall 1974 when a significantly tougher grading scale was instituted (See Figure 4.1, p. 204 and Table 1.2, p. 28.) The present grading scale accurately reflects the student's knowledge of chemistry. The slight inflationary tendency of the retake procedure for unit exams is countered by the deflationary effect of the final exam. The grading scale has been 'fine-tuned' so that a specific grade indicates a level of mastery consonant with faculty expectations and past student performance. The reliability and representativeness of the tests used in these courses are suitable to their purpose.

Student attitudes toward these courses vary in a manner somewhat similar to student achievement. CEM 130 is generally rated

more favorably than is CEM 131 although the average ratings of both are typically near the scale midpoint. The ratings may be lower than those given to other university courses because both percent survey return and course grade correlate positively with favorability of response, and both are lower than for many other courses.

The most popular aspects of the new method are the opportunity to retake examinations, the way the final exam is weighted, and the preset grading scale. The overall mean response on evaluation items for these three aspects is 3.84 on a scale from 1.0 (dislike) to 5.0 (like). The mean is negatively distorted by the five to ten percent of the students with negative attitudes toward the whole system who downgrade every aspect in a 'negative halo' effect.

Chapter V Summary

V.A. Major conclusions

This summary is a brief recapitulation of only the most important and interesting fruits of the research discussed in previous chapters. The complete discussion of results and conclusions for each topic occurs in the various sections denoted in the contents. Observations and conclusions are presented in six areas: (1) test reliability from Chapter II, (2) test validity from Chapters II and IV, (3) survey construction from Chapter III, (4) student achievement and (5) student attitudes, both from Chapter IV, and (6) instructional costs from Chapter I. The seventh section is an overall evaluation of CEM 130/131 under the new instructional system.

V.A.1. Improving test reliability

The surest way to increase test reliability without increasing the number of questions is to lower guessing opportunity. The guessing error observed from four- and five-choice test questions is more than halved by increasing the number of answer choices to ten. Consequently, the test constructor need not sacrifice the convenience of the multiple-choice format in order to write an examination of high reliability and respectably low guessing error.

The reliability or precision of a chemistry test depends on the content of the test questions. Problems are significantly more reliable than nonproblems because testwiseness and other irrelevant skills are

less likely to point to the correct answer when all choices are numbers. However, nonproblems should not be omitted due to their less reliable character. On the contrary, the number of nonproblems should be increased -- even if the test must be made longer -- because simply increasing the number of any type of item increases the overall precision of the total score. The quantity of each format of test question on an examination should be inversely proportional to its reliability.

V.A.2. Ensuring test validity

The accuracy of a chemistry test score is a combination of both absolute and relative validity. Different forms of the same examination should be equivalent in both content and difficulty so that scores on various forms are equally accurate in estimating a student's knowledge. Multiple-choice test questions with eight or ten choices appear to produce tests with the most predictable and equivalent difficulties whereas short-answer items seem least accurate.

Whenever the range of material covered on an exam is larger than the number of questions, the equivalence or parallelism between different forms cannot be perfect unless some material is consistently slighted. Thus once the optimal level of equivalence is reached, there may be no way to decrease such relative inaccuracy without simultaneously increasing absolute inaccuracy.

Absolute accuracy is a result of the procedures used to construct examinations. Exams with statistical equivalence and absolute

accuracy may be constructed most easily by random selections from a pool of every possible question that could be asked. If the material covered by an exam consists of homogeneous sections or strata, random selection should be stratified among these subdivisions.

V.A.3. Attitude survey construction

A direct comparison of two survey questions with these wordings: 'I like something.' [Agree....disagree] and 'How do you feel about something?' [Like....dislike] demonstrated that the positively stated item is not more biased than the neutrally stated item. The only consistent significant bias observed was one item for which the positive format was less -- not more -- biased. Students might construe that they must respond negatively to indicate a neutral attitude -- thus at least in one instance biasing the results in that direction.

The location of the neutral response at other than the scale midpoint is also demonstrated by survey questions of the type 'I would choose A over B'. Twelve comparisons of parallel inversions (two questions and six administrations) produced nine negative midpoints, two neutral midpoints, and one positive midpoint.

These results illustrate the subtle and sometimes unexpected dependence which survey results have on the exact wording of questions. Unless the absolute group attitude has been determined through the laborious procedure of intensity analysis and scaling, results of differently worded surveys should not be directly compared. One must

be limited to relative comparisons between groups which have responded to identical attitude surveys or evaluation forms.

V.A.4. Student achievement in CEM 130/131

Student achievement measured by test scores shows a stable annual pattern. Fall terms of CEM 130 have the highest means, and the achievement in CEM 130 is always higher than in the succeeding term of CEM 131 (for enrollments greater than 300). Course grades are based on the last-tries mean of five or six tests during the term, and students tend to retake tests sufficiently often to maintain the annual achievement pattern.

Under the traditional instructional method, grades in CEM 130 tended to be lower than those in CEM 131. In these courses, the grades in CEM 131 are only slightly lower than under the lecture-recitation method, but the grades in CEM 130 are significantly higher than previously. Grades in CEM 130 now tend to be higher than grades in CEM 131 whereas before, grades in CEM 131 were higher. This new order of mean grades better reflects the difficulty of the material in the two courses.

V.A.5. Student attitudes in CEM 130/131

A surprising relationship between rate of survey return and favorability of response was discovered. Superimposed on the variation in group attitudes from term to term, it was discovered that the mean

response was generally more favorable when the percentage of the group responding to the surveys was higher. The correlation between response and return is not overwhelming, but it is large enough to decrease the significance of small differences between groups with disparate rates of return. It also suggests that the mean responses ought to be considered lower bounds to the total group attitudes.

The overall attitudes of the students toward various aspects of the instructional system are neutral to moderately positive. The most popular aspect of these courses is the testing and grading procedure. Over ninety percent of all students would prefer courses to be taught by methods which include the examination system used in CEM 130/131.

V.A.6. Relative costs of the system

The cost-per-student of a modular self-paced instructional system depends more on course enrollment than does that of a lecture-recitation course. The higher fixed costs in staff and equipment are somewhat offset by lower faculty costs, but the dollar efficiencies of the new method are evident only with large numbers of students. Since the combined enrollments of CEM 130 and CEM 131 average more than fifteen hundred per term excluding summer, the new method has a very slightly lower per-student cost.

An additional efficiency under this system is the cost per contact hour for graduate and undergraduate teaching assistants. The typical assignment for a teaching assistant is twelve contact hours in these

courses. In other courses taught in the Department of Chemistry, the number of assigned contact hours range from three to five for recitation sections in lecture courses. Thus for approximately the same overall dollar cost and student-teacher ratio, the effective contact between students and teaching assistants is more than doubled within these modular self-paced courses.

V.A.7. Evaluation of CEM 130/131

A course may be evaluated in comparison with other courses and judged as relatively more or less successful in student achievement, student satisfaction, and system costs. A course may also be evaluated by its own internal criteria -- whether the proposed goals of the course were met. Relative evaluations are hazardous when made between courses with different types of students, different measuring instruments, and different course content. Yet it does seem true that there are no glaring minuses or dramatic pluses with the new system. Students learn roughly the same chemistry, are approximately as happy with the experience, and cost the department about the same amount of money.

Judging these courses against their own goals provides a slightly different perspective. The goal of student achievement is not explicitly stated, but it is implied that students should certainly learn at least as much as under the lecture-recitation method. They do. (In fact, it is the opinion of some faculty not associated with these courses that

CEM 130 and 131 are more difficult now than in the past.) The goals in student attitude are more ambitious, and perhaps therefore less attainable. It is hoped that all students will find the individualization of subject matter and teaching methods a more enriching experience. Although most students have found these courses enjoyable, a tiny minority are extremely disenchanted with the method and extremely vocal in their dissent. The last goal is the cost of the new system.

One of the major assumptions upon which this new method is based is that instructional personnel must be used more efficiently while the total costs are no greater than before. Now that a semblance of an operational steady state has been reached, the costs are about the same as those for the lecture-recitation method. In addition, the total number of instructional contact hours has increased with no increase in the number of faculty or graduate teaching assistants assigned to these courses. Thus the cost goals set forth in the initial proposal for the new method have been realized.

The new modular self-paced method has been a not unqualified but nevertheless definite success. There is room for minor improvement in system costs, important improvement in test quality and measurement precision, and substantial improvement in student attitudes. These improvements will come with the continuing developmental work of faculty in these courses beyond the routine of course operation.

V.B. Suggestions for future research

As is the case with any project, the questions which have been answered by this research are exceeded by the number of new and different questions which have arisen. The following sections are brief abstracts of proposed research in three areas: achievement, attitude, and prediction of course performance.

V.B.1. In measuring achievement

The typical method of devising wrong answer choices for problems on a multiple-choice test is the 'probable errors' approach. Each wrong choice is the result of some common or easily made mistake in setting up or solving a mathematical expression. It would be much easier to generate eight or ten responses to each question if one did not have to solve a problem through various incorrect routes. Perhaps simply including apparently reasonable 'random' numbers would provide sufficient distraction from the right answer so that students could not answer the question without working the problem.

A parallel situation exists for nonproblems. Some nonnumerical questions simply do not have more than three or four obviously possible wrong answers. The effect on item reliability and difficulty of including totally irrelevant choices -- a kind of random noise -- may not be detrimental. Students who know the answer should not be distracted by the extraneous choices. Conversely, students who do not know the answer to the question or are unsure of the correct

response may become more confused and less able to guess the correct answer.

The division of test questions in this study was into only two classes -- problems and nonproblems. The taxonomies of Bloom and Piaget present several stages of mental development or kinds of thinking. It may be possible to divide test questions into more than two categories based on the thought processes their solutions entail. VanKempen [131] devised a classification scheme for examination questions from rational analysis of the kinds of thinking required to answer them. The observed correlation coefficients between items in the same category were only moderately higher than the correlations between items in different categories.

Instead of defining categories and testing their goodness by contrasting intraclass correlations with interclass correlations, a more fruitful approach might be factor analysis. The relationships which exist between questions are described by the factor structure of the sample. Once the items which cluster are identified, the factors need only be described and labeled -- they already meet the correlation qualifications for 'good' categories. (It might be found that what some test questions have in common are irrelevant characteristics such as position on the test, arrangement of answer choices or sheer length.)

A second dimension along which test questions might be factored is topic or content. Questions might interrelate more strongly when they are from the same topic or unit of material than when they are

from different areas. Thus questions may show type-of-content similarities in addition to type-of-thinking similarities. It would be expected that when both content and thinking are similar, item intercorrelations would be highest, and when both are different, item intercorrelations be lowest.

The different classification schemes -- content and thinking -- have different purposes. A classification of test questions according to kinds of thinking serves a teaching rather than testing function. A student may repeatedly fail test questions which have a specific kind of thinking in common. In this case what is needed is not remediation in a particular unit or area of content, but assistance in learning a cognitive strategy which would apply to any area of content.

The classification according to question content would be used to improve the parallelism between different forms of the same exam. The success of stratified random sampling depends on the internal consistency of the divisions made in the subject matter. When the intercorrelations of items within each division are significantly higher than the correlations between items in different strata, then the reliability and equivalence of parallel test forms can be very high.

Another aspect to be considered when multiple forms of the same test are used is the overall test difficulty. Although it is unrealistic to expect that there be item-by-item correspondence in difficulty between two parallel tests, both tests should be equally difficult as a whole. If tests vary considerably in difficulty, then the use of the same grading

scale for all of them is unjustified. Preliminary indications from this research are that multiple-choice questions with eight and ten answer choices produce the most similar exam difficulties, and that short-answer questions are least similar in difficulty. Considerably more research needs to be done to verify these tentative conclusions and elucidate a possible explanation of the difference in predictability and stability of item difficulty.

V.b.2. In measuring attitudes

The exact wording of a survey question has such an important effect on the responses to it that only identically worded surveys can be compared directly. The importance of wording also extends to the answer choices. Slight changes in a response key may greatly affect the mean or spread of responses. The spacing of response categories is statistically important; if categories are unequally spaced, then most statistical manipulations of the raw numerical data may be unjustified -- even calculation of the mean. The spacing between various categories also depends somewhat on the content of the items.

Answer keys used today are based more on the logic of what seems reasonable rather than on the results of scientific study. Scales with unambiguous and equally spaced intervals need to be established for the kinds of questions asked on attitude and evaluation surveys. Scales with spacings which are too narrow make it difficult for a student to choose a response category; scales with unequal spacings

make it difficult to simplify group responses by mathematical calculations based on the numerical qualities of non-numbers.

A second major influence on survey response is the percent return. The calculations performed in Chapter IV for one of the attitude factors showed that rate of return has a significant effect on response mean. Much work need to be done to determine if this effect is general for all types of survey questions. If specific equations could be devised to correct for the variations in rate of response to diverse questions or types of questions, then surveys with different rates of return could be compared with more confidence. This would allow a better record to be kept of the changes in student attitudes from one term to the next.

V.B.3. In prediction of course performance

Student achievement in CEM 130/131 is the result of a complex interaction of abilities, attitudes, past achievements, expectations, and current activities of the student in the course. It seems reasonable to expect that some of these variables might be used to predict to a certain extent how likely a student is to do poorly or well in a course. If there were some method of identifying at the start of the term those students who will do poorly, perhaps they could be given remediation, special assistance, or academic advising before their poor performance is a matter of permanent record. The percentage of students who enroll in CEM 130/131 and ultimately

receive a grade of 0.0 ranges from three to thirteen percent each term; thus such students are not an inconsequentially small fraction of the entire group.

Some preliminary work in prediction of course performance was attempted. The precourse attitude survey items and several aptitude and achievement scores were variously combined in an attempt to discriminate the poor performer from the rest of the group. The effort was not successful, but it was instructive. Although individual items from the precourse attitude survey do not correlate significantly with course grade, discriminate function analysis of combinations of variables did suggest that differences exist between students ranked according to their course grade.

The precourse attitude survey was not designed for the prediction of course grades. Very few of the questions specifically measure motivations and other attitudes which might affect achievement. The proper experimental approach to creating a course prediction survey is an iterative process of compiling attitude questions and testing their predictive capabilities singly and in combination. Along with such an attitude survey, two other measures provide complementary information. The entrance examination scores of a student, especially on the mathematics tests, indicate abilities, and the score on the first of the six tests indicates course performance. These measures may be combined in a three stage process of 'flunk-risk' identification.

The first stage identifies students likely to do poorly based solely on their high school background, standardized test scores, and entrance examination performance. Not many students would be identified at this stage since such scores typically correlate with course grades in CEM 130 only 0.15 to 0.25. Yet there are students whose math skills or reading ability are so deficient that any college course will be difficult.

The second stage combines the results of a precourse attitude survey with the entrance variables. Students with negative attitudes and expectations as well as low entrance scores are more likely to fail than students notably below average in only one of these areas. The very rough results of the prediction analysis done thus far do suggest a difference in attitude between students at the low end of the grade scale. Students with similar abilities divide into those who scrape by and those who fail completely.

The third and final stage of selection is made after the results of the first exam given in the course has been processed. The interaction of attitude and ability will not produce identical results in all students. Some students may overcome their dislike or fear or underpreparedness to perform adequately. Other students on the edge of predicted failure may need special assistance. This final stage of identification uses the student's actual performance on the first section of course material as a predictor of his achievement on later material.

It might be possible with this stepwise procedure to say after the second week of the term, "Within the total enrollment of 1600 students, this subgroup of 180 contains ninety percent of the 100 students expected to flunk CEM 130." Thus the task of identifying students who might experience difficulty has been shrunk from trying to find the six percent of 1600 to the fifty percent of 180. The selection of students for such programs as Tutorial Assistance in Chemistry might be extended to such high risk students if they could be identified.

REFERENCES

REFERENCES

1. Robert J. Havighurst and Rufus D. Reed, Journal of Chemical Education, v. 18, p. 475 (1941)

2. Lyman C. Newall, Journal of Chemical Education, v. 9, p. 677 (1932)

3. C. A. Browne, Journal of Chemical Education, v. 9, p. 696 (1932)

4. Harrison Hale, Journal of Chemical Education, v. 9, p. 729 (1932)

5. F. B. Davis, Journal of Chemical Education, v. 9, p. 745 (1932)

6. B. Clifford Hendricks, Journal of Chemical Education, v. 2, p. 1187 (1925)

7. William F. Ehret, Journal of Chemical Education, v. 7, p. 321 (1930)

8. S. N. Postlethwaite, J. Novak, and H. T. Murray, Jr., The Audio-Tutorial Approach to Learning, Minneapolis, Burgess, 1969

9. S. N. Postlethwaite, The American Biology Teacher, v. 32, p. 32 (1970)

10. S. N. Postlethwaite and Robert N. Hurst, Educational Technology, v. 12 (No. 9), p. 35 (1972)

11. W. Robert Barnard, E. F. Bertaut, and Rod O'Connor, Journal of Chemical Education, v. 45, p. 617 (1968)

12. W. Robert Barnard, Journal of Chemical Education, v. 45, p. 681 (1968)

13. W. Robert Barnard and Rod O'Connor, Journal of Chemical Education, v. 45, p. 745 (1968)

14. Jesse H. Day, Journal of Chemical Education, v. 39, p. 50 (1962)

15. Jay A. Young, Journal of Chemical Education, v. 40, p. 11 (1963)

16. Jesse H. Day, Journal of Chemical Education, v. 40, p. 14 (1963)

17. Jesse H. Day, Journal of Chemical Education, v. 36, p. 591 (1959)

18. B. F. Skinner, Science, v. 128, p. 969 (1958)

19. S. L. Pressey, School and Society, v. 23, p. 373 (1926)

20. S. L. Pressey, School and Society, v. 36, p. 668 (1932)

21. B. F. Skinner, American Scientist, v. 45, p. 343 (1957)

22. S. Castleberry and J. J. Lagowski, Journal of Chemical Education, v. 47, p. 91 (1970)

23. Carl S. Ewig, J. Thomas Gerig, and David O. Harris, Journal of Chemical Education, v. 47, p. 97 (1970)

24. Stephen K. Lower, Journal of Chemical Education, v. 47, p. 143 (1970)

25. Robert C. Grandey, Journal of Chemical Education, v. 48, p. 791 (1971)

26. Fred S. Keller, Journal of Applied Behavior Analysis, v. 1, p. 79 (1968)

27. Ben A. Green, Jr., Journal of College Science Teaching, v. 1, p. 50 (1971)

28. H. H. Whetzel, The Journal of Higher Education, v. 1, p. 125 (1930)

29. R. H. Schwendeman, Journal of Chemical Education, v. 45, p. 129 (1968)

30. Robert N. Hammer, A Proposal to the General Chemistry Committee and to the Educational Policies Committee for A Modular Multipath General Chemistry Program, Unpublished report, Michigan State University, April, 1972

31. Herbert A. Laitinen, Chemical Analysis, New York, McGraw-Hill, 1960

32. Frederic M. Lord, Educational and Psychological Measurement, v. 17, p. 510 (1957)

33. Frederic M. Lord, Educational and Psychological Measurement, v. 15, p. 325 (1955)

34. Frederic M. Lord, Educational and Psychological Measurement, v. 19, p. 233 (1959)

35. C. Spearman, British Journal of Psychology, v. 3, p. 271 (1910)

36. David Magnusson, Test Theory, Reading, Mass., Addison-Wesley, 1967

37. Julian C. Stanley, Reliability, Ch. 13, Educational Measurement, Robert L. Thorndike, Ed., Washington, D. C., American Council on Education, 1971

38. G. F. Kuder and M. W. Richardson, Psychometrika, v. 2, p. 151 (1937)

39. Walter Kristof, Psychometrika, v. 34, p. 489 (1969)

40. Edward E. Cureton, Educational and Psychological Measurement, v. 18, p. 715 (1958)

41. Frederic M. Lord and Melvin R. Novick, Statistical Theories of Mental Test Scores, Reading, Mass., Addison-Wesley, 1968

42. Frederic M. Lord, Educational and Psychological Measurement, v. 13, p. 517 (1953)

43. Lee J. Cronbach, Psychometrika, v. 16, p. 297 (1951)

44. Edward E. Cureton, Educational and Psychological Measurement, v. 25, p. 327 (1965)

45. K. T. Ng, Educational and Psychological Measurement, v. 34, p. 487 (1974)

46. Robert L. Ebel, Educational and Psychological Measurement, v. 32, p. 249 (1972)

47. H. H. Remmers, Ruth Karslake, and N. L. Gage, The Journal of Educational Psychology, v. 31, p. 583 (1940)

48. H. R. Denney and H. H. Remmers, The Journal of Educational Psychology, v. 31, p. 699 (1940)

49. H. H. Remmers and Edwin Ewart, The Journal of Educational Psychology, v. 32, p. 61 (1941)

50. H. H. Remmers and J. Milton House, The Journal of Educational Psychology, v. 32, p. 372 (1941)

51. H. H. Remmers and H. W. Sageser, The Journal of Educational Psychology, v. 32, p. 445 (1941)

52. H. H. Remmers and R. M. Adkins, The Journal of Educational Psychology, v. 33, p. 385 (1942)

53. Dale Mattson, Educational and Psychological Measurement, v. 25, p. 727 (1965)

54. Amos Tversky, Journal of Mathematical Psychology, v. 1, p. 386 (1964)

55. J. Brown Grier, Journal of Educational Measurement, v. 12, p. 109 (1975)

56. Robert L. Ebel, Educational and Psychological Measurement, v. 29, p. 565 (1969)

57. Frederic M. Lord, Private communication with Robert L. Ebel, June, 1975

58. Frederic M. Lord, Psychometrika, v. 17, p. 181 (1952)

59. Frank Costin, Educational and Psychological Measurement, v. 30, p. 353 (1970)

60. Robert A. Ramos and June Stern, Journal of Educational Measurement, v. 10, p. 305 (1973)

61. Donald W. Zimmerman and Richard H. Williams, Psychological Reports, v. 16, p. 1193 (1965)

62. M. A. Coulter, Psychologia Africana, v. 15, p. 53 (1973)

63. Frederic M. Lord, Psychometrika, v. 20, p. 1 (1955)

64. Harold Gulliksen, Psychometrika, v. 10, p. 79 (1945)

65. Hubert E. Brogden, Psychometrika, v. 11, p. 197 (1946)

66. Lee J. Cronbach and Willard G. Warrington, Psychometrika, v. 17, p. 127 (1952)

67. John B. Carroll, Psychometrika, v. 10, p. 1 (1945)

68. Jum C. Nunnally, Psychometric Theory, New York, McGraw-Hill, 1967

69. William E. Scott, Educational and Psychological Measurement, v. 32, p. 725 (1972)

70. John L. Horn, Educational and Psychological Measurement, v. 31, p. 57 (1971)

71. David M. Shoemaker, Educational and Psychological Measurement, v. 32, p. 705 (1972)

72. Lynnette B. Plumlee, Psychometrika, v. 17, p. 69 (1952)

73. Edward E. Cureton, Educational and Psychological Measurement, v. 31, p. 45 (1971)

74. Louis Guttman, American Sociological Review, v. 12, p. 57 (1947)

75. Allen L. Edwards and Franklin P. Kilpatrick, Psychometrika, v. 13, p. 99 (1948)

76. L. L. Thurstone and E. J. Chave, American Journal of Sociology, v. 33, p. 529 (1928)

77. Rensis Likert, Archives of Psychology, v. 22 (No. 140), p. 1 (1932)

78. Lauren H. Seiler and Richard L. Hough, Empirical Comparisons of the Thurstone and Likert Techniques, Ch. 8, Attitude Measurement, Gene. F. Summers, Ed., Chicago, Rand McNally, 1970

79. Marvin E. Shaw and Jack M. Wright, Scales for the Measurement of Attitude, New York, McGraw-Hill, 1967

80. Lee J. Cronbach, Educational and Psychological Measurement, v. 6, p. 475 (1946)

81. J. P. Guilford, Psychometric Methods, New York, McGraw-Hill, 1954

82. Leonard G. Rorer and Lewis R. Goldberg, Journal of Applied Psychology, v. 49, p. 422 (1965)

83. Leonard G. Rorer, Psychological Bulletin, v. 63, p. 129 (1965)

84. Lois L. Elliot, Educational and Psychological Measurement, v. 21, p. 405 (1961)

85. Bernard DuBois and John A. Burns, Educational and Psychological Measurement, v. 35, p. 869 (1975)

86. Hadley Cantril, Gauging Public Opinion, Princeton, Princeton University Press, 1947

87. Herbert H. Blumberg, Clinton B. DeSoto, and James L. Kuethe, Personnel Psychology, v. 19, p. 243 (1966)

88. C. O. Mathews, The Journal of Educational Psychology, v. 20, p. 128 (1929)

89. Lee J. Cronbach, Educational and Psychological Measurement, v. 10, p. 3 (1950)

90. Amiel T. Sharon, Journal of Applied Psychology, v. 54, p. 278 (1970)

91. David H. Smith, Public Opinion Quarterly, v. 31, p. 87 (1967)

92. Lee Stockford and H. W. Bissell, Personnel, v. 26 (No. 2), p. 94 (1949)

93. Roy C. Bryan, The School Review, v. 52, p. 285 (1944)

94. Barbara H. Showers, Alternative Response Definitions in Instructional Rating Scales, Unpublished Ph. D. dissertation, Michigan State University, East Lansing, 1973

95. Michael S. Matell and Jacob Jacoby, Educational and Psychological Measurement, v. 31, p. 657 (1971)

96. Jacob Jacoby and Michael S. Matell, Journal of Marketing Research, v. 8, p. 495 (1971)

97. Bernard M. Bass, Wayne F. Cascio, and Edward J. O'Connor, Journal of Applied Psychology, v. 59, p. 313 (1974)

98. Karl C. Pratt, Journal of Genetic Psychology, v. 72, p. 201 (1948)

99. D. A. Brotherton, J. M. Reed, and K. C. Pratt, Journal of Genetic Psychology, v. 73, p. 209 (1948)

100. Susan Pepper and Lumbomir S. Prytulak, Journal of Research in Personality, v. 8, p. 95 (1974)

101. Milton D. Hakel, American Psychologist, v. 23, p. 533 (1968)

102. Ray H. Simpson, Quarterly Journal of Speech, v. 30, p. 328 (1944)

103. Robert Strahan and Kathleen C. Gerbasi, Journal of Psychology, v. 85, p. 109 (1973)

104. Chester Schriesheim and Janet Schriesheim, Educational and Psychological Measurement, v. 34, p. 877 (1974)

105. Paul E. Spector, Journal of Applied Psychology, v. 61, p. 374 (1976)

106. Robyn M. Dawes, Fundamentals of Attitude Measurement, New York, John Wiley & Sons, 1972

107. Edward A. Holdaway, The Journal of Experimental Education, v. 40, p. 57 (1971)

108. Allen L. Edwards, Techniques of Attitude Scale Construction, New York, Appleton-Century-Crofts, 1957

109. J. P. Guilford and Ada P. Jorgensen, Journal of Experimental Psychology, v. 22, p. 43 (1938)

110. Erwin K. Taylor and Robert J. Wherry, Personnel Psychology, v. 4, p. 39 (1951)

111. C. I. Hoveland and M. Sherif, Journal of Social and Abnormal Psychology, v. 47, p. 822 (1952)

112. H. H. Kelley, C. I. Hoveland, M. Schwartz, and R. P. Abelson, Journal of Social Psychology, v. 42, p. 147 (1955)

113. Leonard V. Gordon, Educational and Psychological Measurement, v. 31, p. 867 (1971)

114. Peter B. Warr and Thomas L. Coffman, British Journal of Social and Clinical Psychology, v. 9, p. 108 (1970)

115. R. Dubin and T. C. Taveggio, The Teaching Learning Paradox, Eugene, University of Oregon Press, 1968

116. G. C. Chu and W. Schramm, Learning from Television, What the Research Says, Stanford, Stanford University Press, 1968

117. P. Suppes and M. Morningstar, Science, v. 166, p. 343 (1969)

118. H. W. Morrison and E. N. Adams, Modern Language Journal, v. 52, p. 279 (1968)

119. L. D. Muller, Journal of Dairy Science, v. 54, p. 146 (1972)

120. W. J. McKeachie, Y. Lin, and W. Mann, American Educational Research Journal, v. 8, p. 435 (1971)

121. Frank Costin, William T. Greenough, and Robert J. Menges, Review of Educational Research, v. 41, p. 511 (1971)

122. Robert C. Brasted and Kenneth O. Doyle, Strengths and Weaknesses in Formal Student Evaluations of Teaching and the Teacher, Presented at Education in Chemistry '72, Mt. Holyoke, Mass., August, 1972

123. B. R. Siebring and M. E. Schaff, Journal of Chemical Education, v. 51, p. 150 (1974)

124. C. D. Cornwell, Journal of Chemical Education, v. 51, p. 155 (1974)

125. H. H. Remmers, A. D. Taylor, and K. E. Kinter, Student Attitudes Toward Basic Freshman College Subjects and Their Relation to Other Variables, Studies in Higher Education 31, Further Studies in Attitudes, Series 2, Lafayette, Purdue University Press, 1974

126. Donald N. Elliot, Characteristics and Relationships of Various Criteria of College and University Teaching, Studies in Higher Education 70, Further Studies in Attitudes, Series 15, Lafayette, Purdue University Press, 1950

127. William M. Bassin, The Journal of Experimental Education, v. 43, p. 16 (1974)

128. Miriam Rodin and Burton Rodin, Science, v. 177, p. 1164 (1972)

129. R. Barker Bausell and Jon Magoon, Educational and Psychological Measurement, v. 32, p. 1013 (1972)

130. Marian R. Meinkoth, The Journal of Experimental Education, v. 40, p. 66 (1971)

131. Gary W. VanKempen, The Development and Evaluation of a Diagnostic System of Remediation for an Auto-Tutorial Course in General College Chemistry, Unpublished Ph. D. dissertation, Michigan State University, East Lansing, 1977

CROSS REFERENCES

McKeachie, 1971 (120) p. 229
Meinkoth, 1971 (130) p. 238
Menges, 1971 (121) p. 229, 238
Morningstar, 1969 (117) p. 213
Murray, 1969 (8) p. 4
Morrison, 1968 (118) p. 213
Muller, 1972 (119) p. 229

Newall, 1932 (2) p. 2
Ng, 1974 (45) p. 54
Novak, 1969 (8) p. 4
Novick, 1968 (41) p. 51
Nunnally, 1967 (68) p. 66, 123

O'Connor, E, 1974 (97) p. 169
O'Connor, R, 1968 (11) p. 5
O'Connor, R, 1968 (13) p. 5

Pepper, 1974 (100) p. 169
Plumlee, 1952 (72) p. 72, 74, 76
Postlethwaite, 1969 (8) p. 4
Postlethwaite, 1970 (9) p. 4
Postlethwaite, 1972 (10) p. 5
Pratt, 1948 (98) p. 169
Pratt, 1948 (99) p. 169
Pressey, 1926 (19) p. 7
Pressey, 1932 (20) p. 7
Prytulak, 1974 (100) p. 169

Ramos, 1973 (60) p. 59
Reed, J, 1948 (99) p. 169
Reed, R, 1941 (1) p. 1, 29
Remmers, 1940 (47) p. 55
Remmers, 1940 (48) p. 55
Remmers, 1941 (49) p. 55, 168
Remmers, 1941 (50) p. 55
Remmers, 1941 (51) p. 55, 168
Remmers, 1942 (52) p. 55
Remmers, 1974 (125) p. 238
Richardson, 1937 (38) p. 51, 52
Rodin, B, 1972 (128) p. 238
Rodin, M, 1972 (128) p. 238
Rorer, 1965 (82) p. 160
Rorer, 1965 (83) p. 160

Sageser, 1941 (51) p. 55, 168
Schaff, 1974 (123) p. 238
Schramm, 1968 (116) p. 213
Schriesheim, 1974 (104) p. 171
Schwartz, 1955 (112) p. 189
Schwendeman, 1968 (29) p. 15
Scott, 1972 (69) p. 67
Seiler, 1970 (78) p. 157
Sharon, 1970 (90) p. 164
Shaw, 1967 (79) p. 158
Sherif, 1952 (111) p. 189
Shoemaker, 1972 (71) p. 71
Showers, 1973 (94) p. 166, 184
Siebring, 1974 (123) p. 238
Simpson, 1944 (102) p. 171
Skinner, 1957 (21) p. 8
Skinner, 1958 (18) p. 7
Smith, 1967 (91) p. 165
Spearman, 1910 (35) p. 47, 52, 54
Spector, 1976 (105) p. 171, 172
Strahan, 1973 (103) p. 171
Stanley, 1971 (37) p. 50
Stern, 1973 (60) p. 59
Stockford, 1949 (92) p. 165, 184
Suppes, 1969 (117) p. 213

Taveggio, 1968 (115) p. 213
Taylor, A, 1974 (125) p. 238
Taylor, E, 1951 (110) p. 184
Thurstone, 1928 (76) p. 155
Tversky, 1964 (54) p. 57

VanKempen, 1977 (131) p. 255

Warr, 1970 (114) p. 189
Warrington, 1952 (66) p. 64
Wherry, 1951 (110) p. 184
Whetzel, 1930 (28) p. 9
Williams, 1965 (61) p. 59, 69
Wright, 1967 (79) p. 158

Young, 1963 (15) p. 7

Zimmerman, 1965 (61) p. 59, 69

APPENDICES

Appendix A Syllabus, unit objectives, and testable concepts

In this appendix are presented three tables, each of which is an inventory of course material organized for the different purposes set forth in Section IV.A. The three tables are briefly described below.

A.1. Syllabus with Unit divisions for CEM 130/131

The syllabus of topics is divided into units of work approximately equivalent to one lecture period plus associated out-of-class assignment. These unit divisions are designated in the syllabus by lines scored between appropriate subtopic entries. The syllabus presented here is that which was used for the period of this study; it underwent substantial rearrangement in the summer of 1976.

A.2. Objectives for syllabus units of CEM 130 and 131

Each study guide contains a statement of the educational goals which the student is expected to reach on completion of particular units. These objectives are stated in the table as they appear in the various study guides.

A.3. Testable concepts for syllabus units of CEM 130 and 131

Testable concepts are categories of test item content. Although some of the originally defined categories were later consolidated and others went unused, the original numbering system appears in the table. Also included in the table are the number of test items per concept.

Table A.1 Syllabus with Unit divisions for CEM 130/131 (1975)

Topic 1. INTRODUCTION

1. Concerns of the chemist
   a. Relationship of chemistry to other fields of learning
   b. What do chemists think about?
2. Working methods of the chemist
   a. Experimentation
   b. Theoretical studies
   c. Basic research or applied chemistry?
3. Tools of the chemist
   a. The chemical laboratory
   b. Chemical instrumentation
   c. Mathematical tools
4. Matter
   a. Definition of a substance; elements and compounds
   b. Solids and fluids
   c. Mixtures and solutions
   d. The conservation of matter and energy
   e. Interconversion of mass and energy
5. The measurement of matter and energy
   a. Qualitative description and quantitative measurement
   b. Mass and weight
   c. Standards of mass; the metric system
   d. Double and single pan balances
   e. The measurement of energy
      (1) Heat and its measurement
      (2) Temperature and its measurement
6. Physical properties
   a. Definition of physical properties
   b. Extensive and intensive properties
   c. Interconversion of mass and energy
   d. Linear measurement
   e. Volume and density
   f. Comparisons of units, precision in measurement, and probability
7. Atoms and the chemical elements
   a. Composition of atoms
      (1) Electrons, protons, neutrons
      (2) The formation of ions
      (3) Atomic numbers
      (4) Relative weights of atoms; atomic weights
      (5) Isotopes
   b. Names and symbols for the elements
   c. The periodic table
      (1) Groups and periods
      (2) Metals and nonmetals of the periodic table
      (3) Regular and transition elements

Measurement with the analytical balance
   a. Types of balances
   b. "Hands on" balance in CEM Room

Topic 2. CHEMICAL PROPERTIES

1. Chemical compounds
   a. The concepts of oxidation state and oxidation number
   b. Molecules
      (1) Neutral
      (2) Charged
   c. How to write formulas; some rules for the use of oxidation numbers
   d. How to name compounds
      (1) Binary compounds
      (2) Complex nonmetal ions
   e. Acids, bases, and salts
2. Chemical reactions and chemical equations
   a. How to read and write equations
   b. A word about balancing equations
3. Prediction of reaction products; reaction classifications
   a. Metathesis reactions
   b. Displacement reactions
   c. Combination and decomposition reactions
   d. Oxidation and reduction; changes in oxidation number
4. The mole and Avogadro's number
5. Computation of molar quantities
   a. Measurements in grams and in moles
   b. The number of particles in a known mass
      (1) Molecules
      (2) Formula units of ionic substances
   c. The use of units in chemical computations

Table A.1 (cont'd)

6. Computations involving formulas
   a. Calculation of percentage composition from a formula
   b. Calculation of empirical formulas from analytical data
   c. The establishment of molecular formulas

7. Calculations involving chemical equations
   a. Information given by a chemical equation
   b. Use of the mole concept in stoichiometric calculations
      (1) Molar changes described by equations
      (2) Conversion of molar quantities to other units of measurement
      (3) Limiting amounts of reagents

How big is a mole?
a. Calculation of the volume and mass of one-mole samples of five common substances or constant-composition mixtures which have everyday occurrence.

Topic 3. CRYSTALLINE STRUCTURE AND WAVE PHENOMENA

1. Properties of solids
   a. Amorphous substances
   b. Crystalline matter
      (1) Definition of a crystal
      (2) Crystal faces and the constancy of interfacial angles
      (3) Isomorphism and polymorphism

2. Symmetry; symmetry elements and operations
   a. Planes, axes, and the center of symmetry
   b. Tabulation of symmetry elements in a cube and other regular solids
   c. Crystal systems; crystallographic axes
   d. Crystal lattices
      (1) The space lattice
      (2) The unit cell
   e. Implications of symmetry

3. The packing of spheres
   a. Equivalent spheres; structure in metals
      (1) Hexagonal closest packing
      (2) Cubic closest packing
   b. Nonequivalent spheres; ionic crystals
      (1) The radius ratio
      (2) Crystallographic coordination numbers for different types of packing

4. Wave motion
   a. Simple harmonic oscillators
   b. Units of wavelength measurement
   c. Stationaty and travelling waves; nodes
   d. The superposition of travelling waves
      (1) Interference
      (2) Reinforcement
   e. Diffraction of waves

5. Electromagnetic radiation
   a. The electromagnetic spectrum and its importance in chemistry
      (1) Visible radiation
      (2) Other parts of the spectrum
   b. Creation of electromagnetic waves
      (1) Radiofrequency radiation
      (2) Infrared radiation
      (3) Visible and ultraviolet radiation
      (4) X-rays
      (5) Gamma radiation
   c. Detection of electromagnetic radiation

6. The interaction of x-rays with crystals
   a. Diffraction of x-rays
   b. Calculations involving the Bragg equation

7. The Debye-Scherrer powder technique
   a. Fundamental principles
   b. Computation of d-spacings from powder patterns
   c. Use of tables of d-spacings for identifications

8. Structure determination; single crystal studies

9. Electron and neutron diffraction

10. Crystal growth and crystal defects

11. Liquid crystals

Crystals and structure determination
a. Film on diffraction of waves and x-ray technique
b. Models and crystal samples

Table A.1 (cont'd)

Topic 4. THE ELECTRONIC STRUCTURE OF ATOMS

1. The quantization of energy
   a. Failure of the wave theory; black body radiators
   b. Planck's theory
      (1) The quantum
      (2) Planck's constant; comparison with Avogadro's number
2. Spectroscopy
   a. Description
   b. Emission spectra
   c. Absorption spectra
3. The Bohr theory

---

4. Particles and waves
   a. The vibration of a string (a linear vibration)
      (1) Boundary conditions
      (2) Qunatization of vibration of a string
      (3) The "line spectrum" of a string
   b. The vibration of a drum (a vibrating surface)
      (1) Boundary conditions
      (2) Quantization of vibration in a surface; circular symmetry
      (3) The "line spectrum" of a drum surface
   c. Wave equations to describe mechanical vibrations
5. A wave picture of electrons in atoms
   a. The line spectrum of hydrogen
   b. Do electrons have wave properties?
      (1) Light -- photons or waves?
      (2) Application of Planck's and Einstein's equations to photons
      (3) de Broglie's wave treatment of the electron
   c. A wave equation for electrons in atoms
      (1) Spherical vibrations
      (2) Boundary conditions
      (3) Quantization of electron vibrations to fit the line spectrum; discrete energy states
      (4) The schrodinger equation
   d. Locations of electrons in atoms; Heisenberg's uncertainty principle
      (1) The ordinary world versus the atomic world
      (2) Momentum-position and energy-time statements
   e. Electron probability densities and wave properties

---

6. The hydrogen atom
   a. Energy levels of electrons; ground and excited states
   b. Solutions to the wave equation; quantum numbers
      (1) Orbitals
      (2) Symbols, names, and significance of quantum numbers
   c. The first orbital of hydrogen
      (1) Cloud picture
      (2) Point probability -- define and plot
      (3) Radial probability -- define and plot
   d. The second orbital of hydrogen
      (1) Cloud picture
      (2) Radial probability plot

Emission spectroscopy
a. Film on emission spectroscopy
b. "Hands on" use of a Bunsen spectroscope with Geissler tube sources
c. "Hands on" experiments with Nichrome wire flame tests

---

7. The quantum numbers
   a. Probability surfaces
   b. The principal quantum number
      (1) Represents main energy levels
      (2) Possible values
   c. The azimuthal quantum number
      (1) Possible values
      (2) Describes orbital shapes and symmetries
         (a) s orbitals; spherical symmetry
         (b) p orbitals; dumbbells
         (c) Higher quantum number orbitals
   d. The magnetic quantum number
      (1) Possible values
      (2) Orientations of p orbitals
   e. The spin quantum number
      (1) Experimental evidence
      (2) Possible values
      (3) Diamagnetism and paramagnetism
   f. Electron configurations

Topic 5. PERIODIC PROPERTIES OF THE ELEMENTS

1. The aufbau principle
2. Construction of a periodic table based on electron configurations
3. The phenomenon of periodicity
   a. History
   b. Significance

Table A.1 (cont'd)

4.	Periodic properties of the elements a. Melting point and heat of fusion b. Boiling point and heat of vaporization c. Density d. Atomic or molecular volume e. Electrical conductivity f. Thermal conductivity
5.	The measurement of electron attraction a. Ionization potential b. Electron affinity c. Electronegativity
6.	Periodicity in electron attraction a. Trends in the periodic table b. Correlations with electronic structures of the elements
7.	A survey of the elements a. The inert gases (1) Electronic characteristics (2) Compounds of the inert gases b. The regular nonmetals c. The regular metals d. Transition metals e. Lanthanides and actinides
Topic 6.	CHEMICAL BONDING
1.	The concept of a bond
2.	Types of forces between atoms a. Electrostatic forces; simple ionic bonds b. Covalent or electron-pair bonds
3.	Application of wave mechanics to molecular structure a. The H_2^+ and H_2 molecules b. Combinations of atomic orbitals c. Bonding and antibonding orbitals
4.	Molecular orbitals in nonpolar diatomic molecules a. Sigma bonds: formation and symmetry b. Pi bonds: formation and symmetry c. Order of energy levels d. Molecular orbital pictures of H_2, He_2, and Li_2
5.	Electron probability density patterns in covalent bonds a. Effect of electron attraction differences on charge dendity patterns b. Polarizibilities of electron clouds c. Bond strength and cloud density d. The transition from nonpolar bonds to ionization
6.	Correlations between bond types and the properties of compounds
7.	Polyatomic molecules and directed bonds; hybridization
8.	Lewis formulas a. How to write electron-dot structures b. Resonance
	Molecular spectroscopy a. Film on infrared spectroscopy
Topic 7.	MOLECULAR STRUCTURE AND STEREOCHEMISTRY
1.	The significance of directed bonds; stereochemistry
2.	Factors influencing molecular stereochemistry a. Atomic and ionic radii (1) Bond lengths (2) Bond angles (3) Coordination number b. Bond character; restrictions in rotation
	c. Intramolecular and intermolecular interactions (1) The effect of electron repulsion on molecular geometry (2) The role of unshared (lone) pairs (3) The effect of multiple bonding on molecular geometry
3.	Bonding orbitals, lone pairs, and hybridization
4.	Kinds of molecular geometry a. Linear b. Triangular c. Tetrahedral d. Octahedral

Table A.1 (cont'd)

	Symmetry in crystals and molecules a. Identify symmetry elements in a selected group of crystal models b. Identify symmetry elements in various kinds of hybrid orbital sets c. Identify symmetry elements in a selected group of stick-and-ball molecular models and identify real substances whose molecules correspond to the models
Topic 8.	HYDROGEN, OXYGEN, AND THEIR COMPOUNDS
1.	Occurrence and general characteristics of the elements
2.	Production of the elements a. Industrial processes b. Some laboratory reactions that produce hydrogen and oxygen
3.	Electronic structures of the H_2 and O_2 molecules
4.	The formation and reactions of hydrogen and oxygen compounds of regular elements a. Emphasis on hydrogen and oxygen compounds of periods 2 and 3 b. Metal hydrides and oxides (1) Formation (2) Bonding
	(3) Reactions with water (a) Gas evolution (b) Formation of acidic or basic solutions
	c. Compounds containing discrete molecules; nonmetal compounds (1) Formation (2) Bonding (3) Trends in the reaction with water
5.	Summary of trends: relationship between electronic structure, bond type, and properties
Topic 9.	STATES OF MATTER AND THEIR TRANSFORMATIONS
1.	Some properties of solids, liquids, and gases
2.	Systems of matter a. Definition of a system b. Simple phase diagrams (1) Definition of a phase (2) Phase diagram of water; areas, lines, and the triple point
3.	Some properties of gases a. Molecular motion (1) The velocities of gas molecules (2) Diffusion of gases; Graham's law (3) Pressure and its measurement b. Temperature and the motions of gas molecules
4.	The concept of an ideal gas
5.	The ideal gas equation; the gas constant, R
6.	Calculations involving the ideal gas equation
7.	Pressure-volume, temperature-volume and pressure-temperature calculations
8.	Simultaneous temperature-pressure changes
9.	Quantity calculations; molecular weight; density
10.	Mixtures of gases; partial pressures
11.	Deviations from ideal gas behavior; accuracy of the gas laws
12.	The condensation of gases; critical temperature and critical pressure
13.	Structure in liquids
14.	Solid-liquid transformations a. Freezing and melting b. The phenomenon of equilibrium; the symbol c. Solid + heat of fusion → liquid d. Exothermic and endothermic processes e. Le Chatelier's principle; application to phase transformations
15.	Cooling curves and undercooling
16.	Liquid-gas transformations a. Vapor pressures of liquids; equilibrium; evaporation b. Factors influencing vapor pressure values c. Liquid + heat of vaporization → gas d. Boiling point and its variation with pressure

Table A.1 (cont'd)

17. Solid-gas transformations
    a. Vapor pressures of solids; the escaping tendency; equilibrium
    b. Factors influencing vapor pressure values
        (1) Intermolecular forces
        (2) Temperature
    c. Sublimation: solid + heat of sublimation → gas
    d. The normal sublimation temperature
    e. The phase diagram of carbon dioxide

Topic 10. CHEMICAL EQUILIBRIA

1. The extent of chemical reactions
2. Reactions which go to completion
    a. A precipitate is formed
    b. A gas is liberated
    c. A weak electrolyte is produced
    d. Oxidation-reduction occurs
3. The nature of chemical equilibrium; rates of reactions
4. The effect of pressure on equilibria
    a. Solids and liquids
    b. gases
5. The equilibrium constant
    a. The equilibrium law and the equilibrium constant
    b. Correlations with Le Chatelier's principle
6. Effects of experimental conditions on the equilibrium yield of a reaction
    a. Effect of concentration in solution
    b. Effect of pressure
    c. Effect of temperature
7. Calculations based on the equilibrium law (other than solution equilibria)
    a. Homogeneous equilibria (gaseous systems)
    b. Heterogeneous equilibria

Topic 11. SOLUTIONS

1. The nature of solutions; definitions; gaseous, liquid, and solid solutions
2. The solution process
    a. Ionic solutes
    b. Covalent solutes
    c. Solutes which exhibit both ionic and covalent properties
        (1) Soaps and detergents; micelles
        (2) Biologically important molecules
3. Factors which affect the solubility of a solute in a solvent
    a. Molecular aspects of the solution process; character of the solute and solvent
        (1) Nonpolar and polar molecules; hydrogen bonding
        (2) Ion solvation
        (3) Size and charge of ions
        (4) Dielectric constant of the solvent
    b. Energetics of solution; temperature and pressure effects on solubilities
4. Some solubility generalizations
5. Methods of expressing the composition of solutions
    a. Percentage composition (by weight)
    b. Molality
    c. Molarity
    d. Normality
6. Colligative properties of solutions
    a. Explanation and definitions
    b. The vapor pressure-temperature plot for water and solutions
        (1) Nonelectrolytes
        (2) Electrolytes
    c. Freezing point depression
    d. Boiling point elevation
    e. Diffusion through semipermeable membranes
        (1) The process; experimental methods
        (2) Dialysis
        (3) Osmosis
7. Transport of biological material

Topic 12. IONIC EQUILIBRIA

1. The electrical conductivities of solutions; specific and equivalent conductance
    a. Nonelectrolytes
    b. Electrolytes; the formation of ions
        (1) Strong electrolytes; salts, certain acids and bases
        (2) Weak electrolytes; reactions of solutes with solvent
2. Classification of ions
    a. Anions and cations
    b. Simple and complex ions
3. How to write ionic equations

Table A.1 (cont'd)

4. Degree of ionization
   a. Percent ionization of a weak electrolyte
   b. Behavior of strong electrolytes
5. The interionic attraction theory (Debye-Huckel)
   a. The ionic atmosphere; ion solvation
   b. Velocities of ion migration
      (1) The drag effect
      (2) The relaxation effect
   c. Apparent degree of ionization of weak and strong electrolytes
6. Definitions of acids and bases
   a. Arrhenius definitions of acid, base, neutralization
   b. Extension to nonaqueous solvents; solvent system definitions
   c. Lewis definitions of acid, base, neutralization
   d. Advantages and disadvantages of the various approaches
7. Equivalent weights and normalities of acids and bases
   a. The meaning of equivalence
   b. Calculation of the equivalent weight
      (1) Acids and bases
      (2) Oxidizing and reducing agents
8. Applications of the concept of equivalence; titration
   a. The normality of a solution
   b. Weights and equivalent weights from normality and volume; calculation of the number of equivalents in a sample
   c. Titration
      (1) Calculations
      (2) Techniques

Topic 13. EQUILIBRIA IN AQUEOUS SOLUTIONS

1. The ionization of water
   a. The ion product constant for water, K_w
   b. Calculation of H^+ and OH^- concentrations in water
2. The pH scale
   a. Definition of pH
   b. The pH of solutions and their acidity or basicity
3. Equilibria in solutions of weak acids or bases
   a. Examples
      (1) Molecular acids and bases
      (2) Ions that behave as acids or bases
      (3) Hydrated metal ions with acidic or basic properties
   b. Amphoterism
   c. The degree of ionization
   d. Calculation of $[H^+]$ and $[OH^-]$ in solutions of weak acids or bases
4. The common ion effect
5. Successive ionizations of polyprotic acids
   a. Estimation of ionization constants (Pauling's rules)
   b. Calculations
6. Hydrolysis
7. Hydrolysis reactions of ions; bond strengths
   a. Anion hydrolysis
   b. Cation hydrolysis
   c. Salts which do not hydrolyze
8. Hydrolysis of covalent compounds
9. Equilibrium calculations in hydrolysis reactions
   a. The hydrolysis constant
   b. Calculation of solution pH
   c. Calculation of percent hydrolysis
10. Buffer solutions
   a. Definitions and significance
   b. Substances which act as buffers
      (1) A weak acid and a salt of the acid
      (2) A weak base and a salt of the base
   c. Calculation of the pH of buffered solutions
11. Titration curves for acid-base reactions
12. Equilibria involving indicators
   a. The indicator equilibrium constant, K_{in}
   b. The pK of an indicator and its significance
   c. Color changes of some important indicators
13. Heterogeneous equilibria
   a. The solubility product
   b. Calculation of solubility from the solubility product
   c. Precipitation reactions
   d. Problems in the dissolving of a precipitate

Table A.2 Objectives for syllabus units of CEM 130 and 131

Unit	Objectives
1	This unit is intended to suggest how chemists go about their work and to define what chemistry is concerned with....You should get a clear conception of the general nature of matter and energy...You should learn several principles of measurement...You should learn to identify physical properties.
2	You should become acquainted with [the analytical] balance — especially the single pan form used in laboratories everywhere — and with the less precise centigram balances commonly used when less accuracy is needed.
3	In Unit 3 you should learn how to determine the number of electrons, protons, and neutrons in an atom from the mass number and atomic number. At the same time you should become acquainted with some characteristics of electrons, protons, and neutrons and you should come to understand the terms [in the syllabus]. Finally, you should learn the names and chemical symbols of the common elements listed in the study guide .
4	You will first learn about the useful concept of oxidation state. You should be able to find the oxidation number of an element when it is in an electrically neutral compound or a charged compound (an ion). You will then learn a few simple rules for naming compounds. You must learn exactly what information a chemical equation tells you, and you should be able to balance equations when you are finished with this unit.
5	You need to be familiar with terms related to the mole concept and be able to use them to solve problems. You should be able to compute a molecular weight from a given formula....you will learn to measure mass in grams or moles, interconvert these two, and calculate the number of particles in a sample.
6	You should be able to calculate percentage composition given a formula and arrive at a formula given percent composition data.
7	You should be able to tell what happens in a reaction on both a molecular and molar level by inspection of the balanced equation. You should also be able to do calculations involving conversion of moles to grams and vice versa. You should learn to find the limiting reagent in a chemical reaction.
8	You should develop a feeling for the size of a mole.

Table A.2 (cont'd)

Unit	Objectives
9, 10	You should become more familiar with the properties of solids, particularly on a molecular scale. You should see that a formula unit is a representation of a continuous three-dimensional highly ordered aggregate. You should learn the various types of spatial arrangements...and should become familiar with ...symmetry operations which allow these molecular patterns to exist.
11	You should learn the meaning of terms used in measuring waves. You should also learn to describe some different types of waves and how these different types of waves interact with each other and with material objects.
12	You should learn some of the properties of electromagnetic radiation and how these properties change with variations in wavelength of the radiation. You should also learn some of the sources of different types of radiation and some means of detecting radiation in different regions of the electromagnetic spectrum.
13	You will see that the wave properties of ...x-rays lead to interactions with crystalline compounds which give data that allow [one] to determine the positions of atoms or ions in a crystal, and hence their interatomic distances. You will see how calculations with the Bragg equation can be adapted for use with a powdered sample.
14, 15	You will learn that other forms of matter...as well as electromagnetic radiation can undergo diffraction. You will also learn more about crystal growth and note that few, if any, crystals are truly perfect, and will have an opportunity to examine some crystal models so that you can visualize better three-dimensional structures.
16	You should know how observations of the characteristics of black body radiators lead to the quantum theory of energy, what the quantum theory is about, how a spectroscope works, how emission and absorption spectroscopy differ, how emission spectroscopy provides information about atomic structure, and how vibration of strings and drums represent useful analogies to the electronic behavior of atoms, especially with respect to line spectra.
17	You should learn that only discrete energy values are allowable in atoms. You should become acquanited with the line spectrum of hydrogen and its implications, and you should develop an understanding of the Heisenberg uncertainty principle. Finally, you should develop some appreciation for the meaning of electron probability density.

Table A.2 (cont'd)

Unit	Objectives
18	You will learn that the wave nature of electrons is described by the Schroedinger equation and how this equation describes the electron and the orbitals it may occupy in the hydrogen atom.
19	You will observe the emissions of several atoms which you will excite by a flame or an electrical spark. You will learn the definitions of emission and absorption spectra and the structure and uses of a spectroscope.
20	You should become familiar with the quantum numbers — how they arise, what values they have, and what they describe about the orbitals with which they are associated.
21	[You will] learn to write the electron configuration of any element according to a set of rules....understand how the periodic table can be constructed by arranging the elements so that those with similar electron configurations are grouped together....know what is meant by the phenomenon of periodicity, and you should know something of the history and significance of the periodic table.
22	You should memorize trends [in physical and chemical properties of the elements] and should try to rationalize them on the basis of an atom's structure and its position in the periodic table. Take special note of the concept of electronegativity.
23	You should take a closer look at ...electron attraction — and see how it correlates with electron configuration. You should learn the names of important sections of the periodic table and of important groups of elements, and you should learn the general characteristics of the electron configurations of the elements which comprise those parts of the table.
24	You should learn to classify chemical bonds as either primarily electrostatic (ionic) or covalent, and you should recognize the characteristic differences between these classifications. You should learn how covalent bonding results from an extension of our wave picture of the electron associated with one atomic nucleus to a representation of molecular structure in which electron density waves are associated with more than one atomic nucleus. Finally, you should develop some understanding of bonding and antibonding orbital.
25	You will learn to distinguish between sigma and pi bonds, and will learn to write molecular orbital descriptions of diatomic molecules.

Table A.2 (cont'd)

Unit	Objectives
26	You should learn that covalent bonds need not share their electrons equally and that as this disparity in sharing increases, so does the ionic character of the bond. You should also learn that ionic character in a bond influences its strength as well as the properties of the molecule which contains the bond. Finally, you should learn that atomic orbitals are usually not ideal for bonding, but hybridize to form new orbitals with improved bonding capabilities.
27	You should learn how to represent molecules and the bonds in them by electron dot (Lewis) formulas. You should also develop an understanding of the phenomenon of resonance.
28	[You will learn] the mechanism of infrared absorption by molecules and the kind of information which can be obtained from an infrared spectrum of a molecule.
29, 30	You should learn a simple yet powerful method which will enable you to predict molecular geometries and which should increase your understanding of the hybridization concept. You should also learn to predict the geometries of various multiply-bonded molecules.
31, 32	You should learn to predict some of the simpler types of molecular geometries. If you are given a compound in which you know the central atom, the number of sigma bonds, and the number of lone pairs of electrons, you should be able to predict the geometry (shape) and symmetry of the compound, the hybridization of the central atom, and the geometry and symmetry of the hybrid orbitals. If you are given a set of hybrid orbitals, you should be able to identify the spatial arrangement of these orbitals and some of the elements of symmetry present in them. The symmetry elements which you should be able to identify are a mirror plane, a center of symmetry, and an axis of rotation.
33 to 36	[You will learn about] the elements hydrogen and oxygen and the compounds they form with representative elements... [also] the periodic trends in physical properties of these oxides and hydrides... [and] to relate types of chemical bonds to chemical and physical properties.

Table A.2 (cont'd)

Unit	Objectives
1 to 3	In these Units you will learn some properties of solids, liquids, and gases. You will learn to define a system. You should develop an appreciation for the structure, or lack of it, in a gas and you should be able to use the ideal gas equation to calculate various properties of gases.
4	You should develop an understanding of how the pressure, volume, and temperature of gases interact. To do this, you will learn to solve a number of problems involving gases. Finally, you will learn how real gases differ from ideal gases.
5 to 7	You will study [interphase] transformations, the energy associated with them, and the factors affecting these changes. You should become more familiar with the concept of equilibrium and its application. Finally, you will learn about melting points, boiling points, sublimation, and phase diagrams.
8, 9	You should learn to distinguish reversible from irreversible reactions and should know various ways reversible reactions can be driven to completion. Qualitatively, you should be able to use LeChatelier's principle to predict the effect on the composition of an equilibrium system which would result from changes in concentration, pressure, and temperature. Quantitatively, you should be able to write equilibrium law expressions for reversible reactions involving gases, liquids, and solids.
10	You will learn to deal with [chemical equilibria and certain factors affecting them] quantitatively by solving various equilibrium problems.
11	[You will learn] some of the language of solutions, and definitions of terms which you shall encounter repeatedly in this and subsequent units.
12	You will look more closely at factors which affect the solution process and the solubility of substances.
13	[You will learn] the different ways in which the composition of a solution may be expressed... [and] the ability to do problems involving these units.
14	[You will learn to do] problems denoting the effect on colligative properties of solution...in particular the elevation of the boiling point or depression of the freezing point of a solvent by the solute.

Table A.2 (cont'd)

Unit	Objectives
15	You will learn to distinguish strong and weak electrolytes and their formation from both ionic and covalent compounds. You should understand how electrolyte solutions conduct electricity and how electrolyte conduction differs from metallic conduction. You should also learn and practice writing ionic equations which more fully describe a solution than do simple chemical equations.
16	[You should learn about] the extent of ionization of strong and weak electrolytes and of the inter-ionic attraction theory which explains some of the electrical properties of solutions of electrolytes. You should be able to use this theory to explain typical ionic solution behavior.
17	You will learn of several different definitions of acids and bases. You should appreciate the differences in these concepts and be aware of the advantages and disadvantages of each.
18	[You will learn about] equivalence and its application to acid-base reactions. You will review the calculation of equivalent weight and normality.
19	You will learn how to calculate the number of equivalents and the normality for an unknown sample which is titrated to an endpoint with standard solution. You will also learn to calculate the equivalent weight of an unknown acid if a weighed sample is titrated.
20, 21	You should learn how water ionizes and how to calculate acidity (H^+ concentration) or basicity (OH^- concentration) of a water solution from the ion product, K_w, for water. You need to learn the meaning of pH and how to calculate the pH of acidic and basic water solutions. You should be able to write equations for the production of acidic or basic solutions from covalent molecular species, from ions which contain ionizable hydrogen or from hydrated metal ions which can react with water to produce acidic solutions. Finally, you should be able to calculate the degree of ionization of a weak acid or base in water solution of known concentration as well as the H^+ ion and OH^- ion centrations in aqueous solutions of weak acids and bases.
22	You should be able to calculate the change in pH of a weak acid or base solution which is caused by addition of a salt which produces one of the ions of the acid or base. You should be able to estimate values for K_2 and K_3 of a triprotic acid if you know the first ionization constant, K_1.

Table A.2 (cont'd)

Unit	Objectives
23, 24	In Unit 24 the concept of hydrolysis, discussed qualitatively in Unit 23, will be studied from a quantitative viewpoint. You should learn how to calculate the hydrolysis constant of an ion which hydrolyzes, and from comparison of the hydrolysis constants of various ions estimate the relative degree of hydrolysis which occurs and its effect on such quantities as the hydroxide ion concentration or the pH. You should be able to calculate the pH or hydrogen ion concentration of a salt solution which has hydrolyzed.
25	You should learn what a buffer is, how one is made, and what its properties are. You should also be able to calculate the pH of a buffer of particular composition, or determine what composition is necessary to make a buffer with a specific pH.
26	You should learn how the pH of an acidic or basic solution changes as it is titrated. You should be able to sketch a titration curve for a strong acid--strong base titration and you should learn how the shape of the curve changes if the acid or base is a weak electrolyte as well as if a polyprotic acid or a polyhydroxy base is used. You should be able to estimate the pH of the titrated solution at the equivalence point. You must know what an indicator is, what role it plays in a titration and how to choose a suitable indicator after you have estimated the pH at the equivalence point.
27	You [will] learn about equilibria involving ionic substances so slightly soluble that they precipitate from solution. You must learn what a solubility product is, how to calculate its value from solubility data, and how to calculate solubility in moles per liter from a solubility product. For slightly soluble ionic substances, you should be able to determine if precipitation occurs and how to dissolve a precipitate once it is formed.

Table A.3 Testable concepts for syllabus units of CEM 130 and 131

Unit	Concept	Description	Items
1	1	law of conservation of matter and energy	3
	4	specific heat of a substance	5
	5	characteristics of mixtures and solutions	3
	6	comparing compounds, elements, mixtures	1
	7	calculations involving density, mass, volume	28
	10	heat transfer	2
2	1	measurement with balances	5
3	1	general location of elements in table	3
	2	composition of atoms	8
	3	isotopes	3
	4	ions	4
	5	subatomic particles	4
4	1	calculation of oxidation number	10
	3	rules for oxidation numbers	5
	4	definition of binary compound	2
	6	nomenclature	8
	7	exothermic and endothermic reactions	5
	8	translation of equations into words	2
5	1	Avagadro's number	2
	2	calculation of moles, grams from particles	13
	4	definition of mole	2
	5	calculate mass or number from moles	6
	6	calculate mass from number of particles	8
	7	calculate mass from moles	9
	8	mass or mole calculation for diatomic gases	12
	9	comparison of STP volumes of gases	5
	10	calculation of STP volume from mass of gas	14
6	1	calculation of weight percent	12
	2	empirical formula determination	12
	3	molecular formula determination	7
7	1	mole ratios from chemical equations	7
	2	calculation of moles or mass of reagents	6
	3	calculations with an excess reagent	12
	4	calculations with a limiting reagent	9
	5	identification of the limiting reagent	9
8	1	how big is a mole	13

Table A.3 (cont'd)

Unit	Concept	Description	Items
9	1	define isomorphism, polymorphism	3
	2	unit cell	3
	3	symmetry elements	12
	4	properties of crystalline, amorphous	5
10	1	hexagonal and cubic closest packing	4
	2	crystal coordination number	8
11	1	definition of wavelength	4
	2	definition and units of frequency	5
	3	definition and calculation of wavenumber	7
	4	units of length	10
	5	calculation of wavelength, frequency, velocity	12
	6	definition of node	5
	7	definitions of interferences, patterns	5
12	2	electromagnetic wave characteristics	11
	3	wave energy and other characteristics	10
	4	trends in the electromagnetic spectrum	10
13	1	x-ray diffraction, Bragg equation	10
	3	calculation of d-spacing	6
	4	calculation of wavelength	5
	5	calculation of Bragg angles	5
	6	x-ray powder diffraction	3
	7	uses of the powder method	5
	8	cleavage of crystalline, amorphous	3
14	1	electron diffraction	5
	2	neutron diffraction	5
	3	liquid crystals	2
	4	impurities and crystal growth	1
15		redundant with Units 11 - 14	
16	1	definition of quantum, photon	4
	2	Planck quantum theory	5
	3	definition of emission, absorption spectra	10
17	1	hydrogen line spectrum	3
	2	wave, particle nature of light	5
	3	photoelectric effect	2
	4	calculation of deBroglie wavelength	5

Table A.3 (cont'd)

Unit	Concept	Description	Items
	5	calculation of photon energy	5
	6	calculation of emission frequency	5
	7	electron as standing wave	3
18	1	meaning of Schroedinger equation	5
	2	definition of an orbital	1
	3	representations of orbitals	2
	4	ground and excited states of hydrogen	3
	5	symbols for quantum numbers	4
19	2	description of emission, absorption spectra	17
	4	qualitative flame tests	8
	5	structure and use of spectroscope	6
20	1	allowed values of quantum numbers	18
	2	quantum number designations	12
	3	orbital characteristics from quantum numbers	26
	4	forbidden values of quantum numbers	14
	5	number of electrons in (sub)level	14
	6	number of orbitals in an energy level	17
21	1	ground state electron configurations	18
	2	filling orbitals, Aufbau principle	15
	3	Hund's rule	15
	4	electron configuration and chemical similarity	14
22	1	definition of electronegativity, electron affinity	6
	2	ionic sizes	1
	3	ionization potential	6
23	1	identification of classes of elements	14
	2	relative electronegativities	20
	3	relative (non)metallic character	17
	5	periodic trends	5
	6	trends in ionization potential	6
	7	group electron configurations	13
24	1	bond definitions, characteristics	14
	3	MO description of diatomic molecules	10
	4	characteristics of molecular orbitals	6
25	1	molecular orbital configurations	14
	2	description of single/double/triple bonds	8
	3	formation of sigma, pi bonds	5

Table A.3 (cont'd)

Unit	Concept	Description	Items
	4	description of sigma, pi bonds	6
	5	bond order from MO considerations	13
	6	unpaired electrons from MO considerations	9
26	1	relative polarity of covalent bonds	9
	2	molecular polarity predictions	14
	3	comparisons of ionic and covalent compounds	15
	4	descriptions of hybridization	15
27	1	electron dot configurations	7
	2	unshared from electron dot	7
	3	valence shell configurations of atoms	9
	4	resonance structures	7
	5	properties of ionic, covalent compared	5
	6	number of double bonds in molecules	4
28	1	radiation for different spectroscopies	7
	2	description of IR spectroscopy	6
29	1	predicting molecular polarity	13
	2	identification of bonding orbitals	10
	3	hybridized orbitals in a compound	10
	4	molecular geometry and unshared pairs	22
	5	hybridization from shared, unshared pairs	9
30	1	examples of hybridization, geometry	9
	2	symmetry elements in hybrid orbitals	13
	4	lone pairs on central atom	10
31		redundant with Units 29, 30	
32	1	identification from a molecular model	20
33	1	properties of hydrogen and oxygen	10
	2	preparation of hydrogen and oxygen	18
	3	types of reactions	15
	4	identification of oxidation and reduction	14
34	1	preparation, description of hydrides	14
	2	preparation, description of oxides	22
35	1	reactions of hydrides	8
	2	reactions of oxides	27

Table A.3 (cont'd)

Unit	Concept	Description	Items
36	2	hydrogen bonding	10
	3	polarity trends in E-O and E-H bonding	9
	4	trends in bonding and properties	10
1	1	characteristics of states of matter	9
	2	definition, identification of phases	4
	3	phase diagrams	8
2	1	definition, measurement of pressure	8
3	1	calculation of pressure, ideal gas law	10
	2	calculation of temperature, ideal gas law	10
	3	calculation of moles, ideal gas law	10
	4	calculation of volume, ideal gas law	9
4	1	calculation PV constant T	10
	2	calculation VT constant P	7
	3	calculation PT constant V	8
	4	calculation PVT	9
	5	calculation partial pressures	7
	6	calculation of gases mixing	8
	7	ideality of gases	13
	8	calculation of gas density	18
	9	calculation of amount of gas	9
	10	calculation of molecular weight of gas	7
5	1	critical T and P predictions	8
	2	characteristics of liquids	6
	3	miscellaneous properties of liquids	10
	4	melting point, heating/cooling curves	5
	5	calculation of heat of melting/freezing	15
	6	two phase LeChatelier's principle	12
6	1	vapor pressure, descriptive	13
	2	calculation of heat, one phase change	14
	3	calculation of heat, any phase changes	7
7	1	solid-gas transformations	2
8	1	reversible/irreversible reactions	11
	2	reactions going to completion	7

Table A.3 (cont'd)

Unit	Concept	Description	Items
	4	characteristics of equilibrium constants	11
	5	units of equilibrium constants	14
9	1	gas equilibrium changing P, n	12
	2	gas equilibrium changing T	12
	3	gas equilibrium changing P, T, n, solid present	13
	4	writing gas equilibrium expressions	22
	6	mixed phase equilibrium expressions	18
	7	relative amounts and equilibrium constant	12
10	1	calculation of K from equilibrium	12
	2	calculation of equilibrium from K	12
	3	calculation of equilibrium from initial and K	16
	5	calculation of equilibrium from nonequilibrium	13
	7	calculation of new equilibrium from old	15
	9	calculation of K from selected data	14
11	1	solutions definitions	17
12	1	factors affecting solubility	15
	2	trends, examples of solubilities	19
	3	comparative solvation of ions	15
13	1	solution composition from related data	20
	2	calculation of molarity	18
	3	calculation of molarity	18
	4	calculation of amounts from molarity, V	21
	5	calculation of equivalent weight	21
	6	calculation of normality	14
	7	calculation of amounts from normality, V	8
14	1	calculation of freezing point depression	19
	2	calculation of boiling point elevation	18
	3	calculation of amounts from colligative data	18
	4	colligative comparisons	15
15	1	electrochemical definitions	17
	2	recognition of electrolyte strengths	20
	3	writing detailed ionic equations	23
	4	writing net ionic equations	16
16	1	calculation of K from percent ionization	17
	2	nonideal electrolyte behavior	13

Table A.3 (cont'd)

Unit	Concept	Description	Items
17	1	identification of types of acid/base	19
	2	identification of conjugate acid/base	19
18	1	calculation of acid/base equivalent weight	18
19	1	calculation of normality	9
	2	calculation of amount to prepare	35
	3	calculations re titrations	22
	4	calculations of various parameters	19
20	1	calculations with K_w	20
	2	calculation of pH from H	20
	3	calculation of pH from OH	20
	4	calculation of H from pH	29
	5	calculation of H or OH from amount	18
	6	qualitative points on the pH scale	18
21	1	calculations of H, OH from K, amounts	19
	2	calculations of degree ionization from data	17
	3	calculations with common ion present	18
22	2	relationships among K_w, pH, H, et al.	14
	3	successive ionization equations	14
23	1	prediction of salt hydrolysis result	28
	2	prediction of salt hydrolysis result	15
24	2	calculation of K_h for salt	14
	3	calculation of pH of salt solution	15
	4	comparative hydrolysis predictions	16
25	1	calculation of pH for buffer	20
	2	calculation of concentration in buffer of pH	20
	3	identification of a buffer mixture	15
26	1	identification of titration curves	18
	2	selection of titration indicator	13
27	1	solubility product expressions	19
	2	calculation of K_{sp} from solubility	19
	3	calculation of solubility from K_{sp}	19
	4	calculation of solubility with common ion	19
	5	prediction of precipitation	19

Appendix B Statistics and significance tests for fifteen-item tests from various comparison schemes

Eighteen tables are presented in this appendix. Summary statistics which appear in Table B.3 are from computer program Zeus, written specifically for these data. Summary statistics in Table B.12 were calculated by hand for three subgroups of administration code A.

In tables which include significance tests, all test statistic values are adjusted so that the critical value of statistical significance for all tests is approximately 1.99. Significant values are not marked inasmuch as many of the significance tests performed on these data are 'liberal' and produce random significant results which may not indicate true differences. Only when important trends in groups of data exist are several statistically significant results accorded recognition. These trends have been discussed in Chapter II.

Table B.1 Glossary of symbols

Symbol	Description or definition
α	alpha level of statistical significance; proportion of studies for which a statistical procedure will lead to false rejection of the null hypothesis
d	difference in test means
df	degrees of freedom for a statistical test of significance
E	error score; number or proportion incorrect
F	F value for an F-ratio test of equality
FP	fraction of a test's questions which are problems
ICS	item chance score; reciprocal of the harmonic mean number of answer choices
J	number of answer choices per item
k	test length; number of test questions
$\widehat{KR}_3$	Kuder-Richardson Formula 3 for test reliability: $r_{tt} = \frac{s_t^2 - \sum pq + \sum \hat{r}_{ii} pq}{s_t^2}$ where $\hat{r}_{ii}$ is estimated by the mean tetrachoric and phi correlations of one item with all other items
KR_8	Kuder-Richardson Formula 8 for test reliability: $r_{tt} = \frac{s_t^2 - \sum pq}{s_t^2} + \sqrt{\frac{\sum r_{it}^2 pq}{s_t^2} + \left(\frac{s_t^2 - \sum pq}{2s_t^2}\right)^2}$ where r_{it} is the biserial item-test correlation
KR_{20}	Kuder-Richardson Formula 20 for test reliability: $r_{tt} = \left(\frac{k}{k-1}\right)\frac{s_t^2 - \sum pq}{s_t^2}$
KR_{21}	Kuder-Richardson Formula 21 for test reliability: $r_{tt} = \left(\frac{k}{k-1}\right)\frac{s_t^2 - k\bar{p}\bar{q}}{s_t^2}$

Table B.1 (cont'd)

Symbol	Description or definition
MCE	multiple choice equivalent number of answer choices; the harmonic mean of the number of choices per item with the assumption that a short-answer item has an infinite number of answer choices
MS_w	mean square within; synonymous with mean test variance
MS_b	mean square between; related to covariance
n, N	number of students in a group or measurements in a collection of data
p	proportion of a group which answers an item correctly
$\bar{p}$	mean proportion correct for a test
Δp	difference in mean proportions correct between tests
P less than	alternate notation for a level of significance similar to the alpha level
q	proportion of a group which answers an item incorrectly
q′	q value adjusted so that the critical q is equal to 1.99 (which is approximately the same as the critical t value)
q value	post hoc comparison of means with the form: $q = \frac{d}{SED}$ where the resulting q value is compared against the critical q value
Q′	q value for correlated samples adjusted so that the critical q is equal to 1.99; differs from the just mentioned q value only in the calculation of a standard error of the difference (SED) which is put into the equation
r	general symbol for reliability or correlation; several of the common subscripted or modified types of r values are described on the next page

Table B.1 (cont'd)

Symbol	Description or definition
r_{ii}	item reliability; operationally undefined but may be estimated from the tetrachoric and phi correlations of an item with all other items
r_{it}	correlation between an item and the test; here the biserial item-test correlation coefficient is used: $r_{bis} = \frac{X_{.1} - X_{.0}}{s_x^2}\left(\frac{n_1 n_0}{u n\sqrt{n^2 - n}}\right)$ where $X_{.1}$ is the mean of those scoring correct on the item, $X_{.0}$ is the mean of those scoring incorrect on the item, n_1 is the number scoring correct, n_0 is the number scoring incorrect, and u is the ordinate (height) of the unit normal distribution at the point above which lies $100(n_1/n)$ percent of the area under the curve
r_{ij}	correlation between one item and another item
r_{tt}	correlation between one test and another test or the reliability of a test
$_k r$	test reliability adjusted by the Spearman-Brown prophecy formula to a standard length of k items; in these pages $_{15}r$ and $_{10}r$ are most often used
r'	reliability coefficient adjusted for guessing error according to the equation: $r' = \frac{2J - 1}{J^2} + \left(\frac{(J - 1)^2}{J^2}\right)\frac{s_t^2 r}{s_t^2 - (k\bar{q}/J)}$
σ	population standard deviation
σ^2	population variance
s	sample standard deviation
s^2	sample variance
Δs^2	difference in variances between tests

Table B.1 (cont'd)

Symbol	Description or definition
S'	S value adjusted so that the critical S is equal to 1.99 (which is approximately the same as the critical t value)
S value	post hoc comparison of means with the form: $S = \frac{d}{SED}$
SED	standard error of the difference; similar in conception to the standard deviation - differences are significant when they are greater than or equal to a certain number of standard errors of the difference
-	t-test, independent samples, equal n's $SED = \sqrt{\frac{s_x^2 + s_y^2}{n}}$
-	t-test, independent samples, unequal n's $SED = \sqrt{\frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{(n_x - 1) + (n_y - 1)}\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}$ It was found that never were variances and sample sizes different enough in practice to yield a different SED than when it is assumed sample sizes are equal.
-	t-test, dependent or correlated samples $SED = \sqrt{\frac{s_x^2 + s_y^2}{n} - \frac{2 r_{xy} s_x s_y}{n}}$
-	q-test, independent samples $SED = \sqrt{\frac{MS_w}{n}}$
-	q-test, dependent samples $SED = \sqrt{\frac{MS_w - \bar{r} MS_w}{n}}$

Table B.1 (cont'd)

Symbol	Description or definition
-	S-test, independent samples, one mean with one mean $SED = \sqrt{\frac{2MS_w}{n}}$
-	S-test, independent samples, complex comparison of the average of three means with the average of three means $SED = \sqrt{\frac{2MS_w}{3n}}$
-	variances t-test, dependent samples; instead of F-ratio $SED = \sqrt{\frac{4 s_x^2 s_y^2}{n-2}(1 - r_{xy})}$
SEM	standard error of the measurement $SEM = s_x \sqrt{1 - r}$ where r is the reliability of test x
SEM	stability estimate of the mean; similar in conception to the standard deviation $SEM = \sqrt{\frac{\sum\left(\frac{d}{2}\right)^2}{n-1}}$
t	t value for a t-test of the significance of a difference $t = \frac{d}{SED}$
t	test or number of tests
VF	variance fraction; proportion of observed variance labeled by a subscript
x, y	scores or variables

Table B.2 Date, source, and number of forms for each code

Code	Date	Source	Forms	Comparisons	
				Unique	Total
A	21 February 1874	CEM 131 W74 E4	3	3	9
B	30 April 1974	CEM 131 S74 E3	2	1	1
C	2 May 1974	CEM 131 S74 E3	4	2	2
D	14 May 1974	CEM 131 S74 E4	2	1	2
E	16 May 1974	CEM 131 S74 E4	4	3	4
F	28 May 1974	CEM 131 S74 E5	2	1	1
G	30 May 1974	CEM 131 S74 E5	4	2	2
H	9 April 1975	CEM 130 S75 E1	4	1	2
I	17 April 1975	CEM 130 S75 E2	2	1	1
J	17 April 1975	CEM 131 S75 E2	4	1	3*
K	8 May 1975	CEM 131 S75 E4	3	1	2
L	25 November 1975	CEM 130 F75 E6	4	1	2
M	26 November 1975	CEM 130 F75 E6	2	1	1*
N	14 January 1976	CEM 131 W76 E1	4	2	3*

*One of the comparisons was for a group too small to be statistically reliable. In this collection of data, three groups -- J3, M, and N3 -- were found to be too small and these data should be lightly regarded.

Table B.3 Summary statistics for parallel forms reliability estimation

Code*	N	r_{tt}**	$\bar{p}$	s^2	KR_8	$\widehat{KR}_3$	KR_{20}	KR_{21}	$\bar{r}_{ij}$	FP	MCE
A1-2	296	.6040	.6491	7.476	.8075	.7114	.6860	.5817	.198	.60	4.7
A1-3	296		.6815	6.255	.7368	.6147	.5905	.5137	.156	.60	4.7
A2-1	284	.5566	.6784	6.881	.7636	.6601	.6288	.5604	.168	.60	4.8
A2-3	284		.6847	6.184	.7324	.6114	.5870	.5104	.155	.60	4.7
A3-1	277	.6556	.6777	6.965	.7751	.6658	.6448	.5675	.169	.60	4.8
A3-2	277		.6366	7.553	.8108	.7178	.6915	.5865	.204	.60	4.7

*The test administration code is designated by the letter to the left of the hyphen. When more than one comparison occurred on the same date, this letter is paired with a number. The form number of the examination is indicated by the number to the right of the hyphen.

**The intercorrelation between two examinations is the geometric mean of their reliability coefficients [44]. From the three possible intercorrelation coefficients and this relationship, a set of simultaneous equations can be solved for the three individual reliability coefficients. When only two tests are compared there are two unknown reliability coefficients and they cannot be determined from only one known intercorrelation.
For this administration the following reliabilities were calculated: $r_1 = .6041$, $r_2 = .7114$, $r_3 = .5128$.

Table B.3 (cont'd)

Code	N	r_{tt}	$\bar{p}$	s^2	KR_8	$\widehat{KR}_3$	KR_{20}	KR_{21}	$\bar{r}_{ij}$	FP	MCE
D1-3	72	.7260	.5926	7.537	.7856	.6801	.6578	.5566	.180	.80	5.7
D1-2	72		.6056	7.937	.8060	.7022	.6799	.5878	.190	.80	5.7
D2-1	76	.7552	.6596	8.469	.8238	.7414	.7157	.6454	.219	.80	5.7
D2-4	76		.5588	7.892	.7865	.6912	.6699	.5694	.177	.80	5.7
E1-5	86	.5602	.5791	8.453	.8007	.7176	.6938	.6081	.208	.87	5.6
E1-2	86		.6705	9.185	.8774	.8013	.7722	.6848	.276	.87	5.3
E2-3	62	.5805	.6441	8.326	.8367	.7491	.7227	.6289	.225	.73	5.3
E2-2	62		.6505	10.809	.8965	.8322	.8025	.7334	.338	.73	5.3
E3-7	73	.6127	.6320	7.559	.8003	.6943	.6677	.5769	.196	.73	5.7
E3-4	73		.6000	9.306	.8160	.7346	.7150	.6570	.202	.73	5.3
E4-1	58	.5858	.6920	10.380	.8574	.7958	.7859	.7360	.228	.73	5.3
E4-4	58		.6046	8.276	.7883	.7094	.6842	.6045	.202	.73	5.3

Table B.3 (cont'd)

Code	N	r_{tt}	$\bar{p}$	s^2	KR_8	$\widehat{KR}_3$	KR_{20}	KR_{21}	$\bar{r}_{ij}$	FP	MCE
K1-3	66	.7609	.5232	10.007	.8239	.7542	.7287	.6708	.233	.73	26.8
K1-2	66		.5737	11.135	.8633	.7994	.7736	.7185	.274	.73	27.7
K2-1	65	.8666	.5456	11.059	.8605	.7978	.7673	.7111	.286	.73	27.7
K2-3	65		.5108	14.571	.9008	.8459	.8255	.7958	.330	.73	26.8
H1-2	40	.8402	.5767	9.669	.8163	.7159	.7071	.6656	.165	.80	36.0
H1-4	40		.6667	12.974	.8977	.8285	.8127	.7961	.279	.80	40.0
H2-1	44	.8618	.6864	9.841	.8582	.7862	.7553	.7199	.281	.80	40.0
H2-3	44		.6273	11.643	.8844	.8228	.7978	.7487	.315	.80	36.0

Table B.3 (cont'd)

Code	N	r_{tt}	$\bar{p}$	s^2	KR_8	$\widehat{KR}_3$	KR_{20}	KR_{21}	$\bar{r}_{ij}$	FP	MCE
L1-1	86	.7374	.5674	8.441	.7451	.6762	.6664	.6020	.153	.00	6.4
L1-3	86		.6124	8.530	.8136	.7255	.7041	.6242	.203	.00	6.2
L2-2	80	.5540	.6267	6.699	.7853	.6742	.6460	.5101	.174	.00	6.2
L2-4	80		.5867	7.175	.7466	.6258	.6092	.5283	.132	.00	6.4
J1-1	109	.5810	.7119	7.387	.8206	.7281	.6969	.6252	.224	.20	7.3
J1-3	109		.7040	4.323	.7355	.5847	.5526	.2966	.158	.20	7.1
J2-2	112	.7717	.6690	5.350	.7708	.6208	.6048	.4063	.108	.20	7.1
J2-4	112		.6625	9.519	.8563	.7782	.7541	.6939	.250	.20	7.3
N1-3	170	.6066	.6329	7.021	.7661	.6667	.6316	.5396	.193	.60	9.1
N1-2	170		.6075	6.242	.7355	.6028	.5843	.4575	.122	.53	8.7
N2-1	153	.6280	.7181	5.612	.7308	.6016	.5716	.4917	.149	.60	8.9
N2-4	153		.6349	7.646	.7884	.6862	.6640	.5842	.181	.60	9.1

Table B.3 (cont'd)

Code	N	r_{tt}	$\bar{p}$	s^2	KR_8	$\widehat{KR}_3$	KR_{20}	KR_{21}	$\bar{r}_{ij}$	FP	MCE
B-2	133	.7237	.5208	7.760	.7777	.6793	.6570	.5545	.165	.40	15.8
B-1	133		.6531	9.587	.8444	.7658	.7430	.6916	.235	.40	4.7
C1-2	144	.6746	.5833	6.916	.7933	.6819	.6603	.5067	.155	.40	15.8
C1-1	144		.7343	5.273	.7283	.6026	.5749	.4767	.142	.40	4.4
C2-2	116	.6538	.5672	7.452	.7990	.7073	.6769	.5420	.211	.47	16.7
C2-1	116		.6477	4.953	.6987	.5640	.5268	.5566	.163	.47	4.6
F-2	74	.7969	.3721	9.397	.8304	.7229	.7098	.6719	.164	.40	19.6
F-1	74		.4559	11.206	.8451	.7728	.7537	.7157	.225	.40	9.2

Table B.3 (cont'd)

Code	N	r_{tt}	$\bar{p}$	s^2	KR_8	$\widehat{KR}_3$	KR_{20}	KR_{21}	$\bar{r}_{ij}$	FP	MCE
G1-2	92	.6856	.3906	10.980	.8687	.7904	.7723	.7157	.241	.47	15.0
G1-1	92		.4246	10.280	.8283	.7455	.7271	.6895	.211	.47	9.8
G2-1	115	.7794	.4296	9.109	.7767	.6892	.6739	.6391	.162	.47	9.8
G2-2	115		.4139	9.939	.8325	.7444	.7267	.6791	.199	.47	15.0
I-1	59	.5855	.6599	4.127	.6250	.4215	.3966	.1975	.077	.13	11.6
I-2	59		.7853	4.520	.7964	.6530	.6101	.4719	.150	.13	11.6
J3-3	26	.4458	.6205	5.102	.7351	.5620	.5526	.3296	.022	.20	7.1
J3-1	26		.6051	5.754	.7055	.5461	.5331	.4041	.092	.20	7.3
M-5	20	.8767	.6067	12.621	.8853	.8111	.7989	.7676	.251	.00	6.3
M-6	20		.6233	12.871	.9267	.8595	.8365	.7783	.364	.00	6.0
N3-2	21	.5373	.6667	6.400	.7888	.6597	.6364	.5134	.162	.53	8.7
N3-1	21		.6698	9.548	.8529	.7483	.7376	.6992	.177	.60	8.9

Table B.4 Comparison of forms by group for dependent samples

Code	N	d	SED	t	q′	Q′
A1-2, A1-3	296	.486	.135	3.57	0.67	1.19
A2-1, A2-3	284	.095	.143	0.66	0.13	0.23
A3-1, A3-2	277	.617	.136	4.57	0.85	1.51
D1-3, D1-2	72	.195	.242	0.80	0.27	0.48
D2-1, D2-4	76	1.512	.230	6.58	2.09	3.71
E1-5, E1-2	86	1.371	.300	4.56	1.90	3.36
E2-3, E2-2	62	.096	.362	0.26	0.13	0.24
E3-7, E3-4	73	.480	.300	1.60	0.67	1.18
E4-1, E4-4	58	1.311	.367	3.58	1.82	3.22
B-2, B-1	133	1.985	.191	10.38	2.75	4.87
C1-2, C1-1	144	2.265	.168	13.52	3.14	5.56
C2-2, C2-1	116	1.208	.196	6.16	1.67	2.96
F-2, F-1	74	1.257	.240	5.25	1.74	3.08

Table B.4 (cont'd)

Code	N	d	SED	t	q′	Q′
K1-3, K1-2	66	.758	.277	2.73	1.05	1.86
K2-1, K2-3	65	.522	.236	2.21	0.72	1.28
H1-2, H1-4	40	1.350	.309	4.37	1.87	3.31
H2-1, H2-3	44	.887	.262	3.36	0.36	2.18
L1-1, L1-3	86	.675	.228	2.97	0.32	1.66
L2-2, L2-4	80	.600	.278	2.16	0.39	1.47
J1-1, J1-3	109	.119	.217	0.55	0.30	0.29
J2-2, J2-4	112	.098	.186	0.53	0.14	0.24
N1-3, N1-2	170	.381	.175	2.17	0.53	0.94
N2-1, N2-4	153	1.248	.181	6.88	1.73	3.06
G1-2, G1-1	92	.510	.270	1.89	0.71	1.25
G2-1, G2-2	115	.236	.191	1.23	0.33	0.58
I-1, I-2	59	1.881	.247	7.63	2.61	4.62
J3-3, J3-1	26	.231	.481	0.48		
M-5, M-6	20	.249	.396	0.63		
N3-2, N3-1	21	.047	.600	0.08		

Table B.5 Comparison of forms by test date for dependent samples

Code	N	d	SED	t	q′	Q′
A-1, A-2	283.5	.528	.136	3.88		1.30
A-1, A-3	285.3	.075	.143	0.52		0.18
A-2, A-3	282.3	.603	.135	4.47		1.48
D-34, D-12	148	.854	.167	5.11	1.18	2.10
E1-5, E1-2	86	1.371	.300	4.57	1.90	3.36
E-34, E-12	120	.704	.258	2.73		1.73
E3-7, E3-4	73	.480	.300	1.60		1.18
B-2, B-1	58	1.311	.367	3.58	1.82	3.22
C1, C2	130	.770	.308	2.50		1.89
F-2, F-1	74	1.257	.240	5.25	1.74	3.08
K-3, K-12	131	.640	.183	3.50		1.57
H-23, H-14	84	1.118	.198	5.64	1.55	2.74
L-14, L-23	166	.638	.181	3.51		1.57
J-14, J-23	221	.011	.145	0.08		0.03
N1-3, N1-2	170	.381	.175	2.17		0.94
N2-1, N2-4	153	1.248	.181	6.88	1.73	3.06
G-2, G-1	207	.373	.162	2.31		0.92
I-1, I-2	59	1.881	.247	7.63	2.61	4.62

Table B.6 Comparison of groups by test date for independent samples

Code	N	d	SED	t	q′	S′
D1, D2	74	.152	.464	0.33		0.18
E2, E4	60	.015	.561	0.03		0.02
C1, C2	130	.770	.308	2.50		0.90
K1, K2	60	.305	.624	0.49		0.36
H1, H2	42	.528	.725	0.85		0.62
L1, L2	83	.252	.443	0.57		0.30
J1, J2	110.5	.633	.347	1.83		0.74
N1, N2*	161.5	.845	.287	2.95		0.99
G1, G2	103.5	.293	.441	0.66		0.34

*Only one of the two tests taken by these groups was the same form.

Table B.7 Comparison of positions by test date for dependent samples

Code*	N	d	SED	t	q′	S′
D.1, D.2	148	.659	.328	2.01	1.39	1.01
E.1, E.2	120	.608	.397	1.53		0.93
K.1, K.2	131	.118	.423	0.28		
H.1, H.2	84	.232	.513	0.45		
L.1, L.2	166	.038	.305	0.12		
J.1, J.2	221	.108	.245	0.44		
G.1, G.2	207	.137	.312	0.44		

*The number after the decimal is the position in which the exam was taken.

Table B.8 Comparison of positions by group for independent samples

Code	N	d	SED	t	q′	S′
D1-3, D2-1	74	1.005	.465	2.16	1.39	0.83
D2-4, D1-2	74	.702	.462	1.51	0.97	
E2-3, E4-1	60	.719	.558	1.29	1.00	
E4-4, E2-2	60	.689	.564	1.22	0.95	
K1-3, K2-1	60	.336	.593	0.57	0.47	
K2-3, K1-2	60	.944	.655	1.44	1.31	
H1-2, H2-1	42	1.646	.682	2.41	2.28	1.36
H1-4, H2-3	42	.591	.766	0.77	0.82	
L1-1, L2-2	83	.890	.427	2.08	1.23	
L2-4, L1-3	83	.386	.435	0.89	0.53	
J1-1, J2-2	110.5	.644	.340	1.89	0.89	
J2-4, J1-3	110.5	.623	.354	1.76	0.86	
N1-3, N2-1	161.5	1.278	.280	4.57	1.77	1.06
N1-2, N2, 4	161.5	.411	.293	1.40	0.57	
G1-2, G2-1	103.5	.585	.441	1.33	0.81	
G2-2, G1-1	103.5	.161	.442	0.26	0.22	

Table B.9 Comparison of groups by form for independent samples

Code	N	d	SED	t	q′	S′
D1-3, D2-4	74	.507	.457	1.11	0.70	
D2-1, D1-2	74	.810	.471	1.72	1.12	
E2-3, E4-4	60	.593	.414	1.43	0.82	
E4-1, E2-2	60	.623	.594	1.05	0.86	
K1-3, K2-3	60	.186	.640	0.29	0.26	
K2-1, K1-2	60	.422	.608	0.69	0.58	
H1-2, H2-3	42	.759	.712	1.07	1.05	
H2-1, H1-4	42	.296	.737	0.40	0.41	
L1-1, L2-4	83	.290	.434	0.67	0.40	
L2-2, L1-3	83	.260	.428	0.50	0.36	
J1-1, J2-4	110.5	.741	.391	1.89	1.03	
J2-2, J1-3	110.5	.525	.296	1.77	0.73	
N1-3, N2-4	161.5	.030	.301	0.01	0.04	
G1-2, G2-2	103.5	.350	.450	0.78	0.48	
G2-1, G1-1	103.5	.075	.433	0.17	0.10	

Table B.10 Comparison of variances by form for dependent samples

Code	df	r	Δs^2	t	P less than
A1-2, 1-3	294	.6040	1.22	2.43	.02
A2-1, 2-3	282	.5566	0.70	1.35	
A3-1, 3-2	275	.6556	0.59	1.15	
E2-3, 2-2	60	.5805	2.48	1.57	
E4-1, 4-4	56	.5858	2.10	1.32	
B-1, B-2	131	.7237	1.83	2.31	.03
C1-2, 1-1	142	.6746	1.64	2.84	.01
C2-2, 2-1	114	.6538	2.50	3.73	.01
F-2, F-1	74	.7969	1.81	1.66	
K2-1, 2-3	63	.8666	3.51	3.01	.01
H1-2, 1-4	38	.8402	3.31	2.28	.03
H2-1, 2-3	42	.8618	1.80	1.47	
J1-1, 1-3	107	.5810	3.06	4.33	.01
J2-2, 2-4	110	.7717	4.17	3.21	.01
N2-1, 2-4	151	.6280	2.03	3.66	.01

Table B.11 Comparison of variances by group for independent samples

Code	df	s^2	s^2	F	P less than
D1, D2	70, 74	26.70	28.71	1.08	
E1, E2	84, 60	27.51	30.15	1.10	
E1, E3	84, 71	27.51	27.14	1.01	
E1, E4	84, 56	27.51	29.52	1.07	
E2, E3	60, 71	30.15	27.14	1.11	
E2, E4	60, 56	30.15	29.52	1.02	
E3, E4	71, 56	27.14	29.52	1.09	
C1, C2	142, 114	20.34	20.35	1.00	
K1, K2	64, 63	37.21	47.63	1.28	.17
H1, H2	38, 42	41.46	39.93	1.04	
L1, L2	84, 78	29.49	21.56	1.37	.09
J1, J2	107, 110	18.28	25.88	1.41	.04
N1, N2	168, 151	21.29	21.49	1.01	

Table B.12 Summary data for the nine subgroups within Code A

Code*	N	$\bar{p}$	s^2	s	r_{it}	KR_{20}	KR_{21}
W1.1	102	.6813	5.91	2.43	.5374	.5735	.4809
W2.2	102	.6720	6.93	2.63	.6015	.6726	.5603
W3.3	102	.7000	6.35	2.52	.5780	.6051	.5399
X2.1	105	.6273	8.00	2.82	.5943	.7028	.6018
X3.2	105	.6620	7.66	2.77	.5728	.6671	.6020
X1.3	105	.6640	7.07	2.66	.5723	.6425	.5643
Y3.1	100	.6820	4.81	2.19	.5072	.4614	.3468
Y1.2	100	.6873	6.20	2.43	.5519	.6248	.5143
Y2.3	100	.6540	6.29	2.50	.5699	.6274	.4932
Z1	307	.678	6.34	2.52	.554	.617	.518
Z2	307	.651	6.84	2.62	.588	.673	.518
Z3	307	.681	6.24	2.50	.553	.591	.511

*The code letter indicates the group; the code number is of the form x.y where x is the form and y is the position. Code Z is the summation across three groups for the same test form regardless of position.

Table B.12 (cont'd)

Code	N	r_{tt}	$\overline{KR}_{20}$	$\Delta\bar{p}$	Δs^2	$\bar{s}^2$	ICS
W1.1, 2.2	102	.6364	.6231	.0093	1.02	6.42	.212
X2.1, 1.3	105	.6510	.6727	.0367	0.93	7.54	.212
Y1.2, 2.3	100	.5890	.6261	.0280	0.09	6.25	.212
Z1, 2	307	.6273	.645	.027	0.50	6.59	.212
W1.1, 3.3	102	.5871	.5893	.0187	0.44	6.13	.212
X3.2, 1.3	105	.6186	.6548	.0020	0.59	7.37	.212
Y3.1, 1.2	100	.5033	.5431	.0053	1.39	5.51	.212
Z1, 3	307	.5731	.604	.003	0.10	6.29	.212
W2.2, 3.3	102	.6133	.6389	.0280	0.58	6.64	.213
X2.1, 3.2	105	.7145	.6850	.0347	0.34	7.83	.213
Y3.1, 2.3	100	.5278	.5444	.0333	1.48	5.55	.213
Z2, 3	307	.6299	.632	.030	0.60	6.54	.213

Table B.13.1 Form and group ANOVA for Codes W, X, and Y

Source	df	MS	F	P less than
Within	912	6.60		
Form	2	18.33	2.78	.06
Group	2	20.16	3.06	.05
Interaction	4	10.58	1.60	.17

Table B.13.2 Group and position ANOVA for Codes W, X, and Y

Source	df	MS	F	P less than
Within	912	6.60		
Group	2	20.16	3.06	.05
Position	2	2.33	0.35	>.25
Interaction	4	10.58	1.60	.17

Table B.13.3 Form and position ANOVA for Codes W, X, and Y

Source	df	MS	F	P less than
Within	912	6.60		
Form	2	18.33	2.78	.06
Position	2	2.33	0.35	>.25
Interaction	4	10.58	1.60	.17

Table B.14.1 Comparison of positions for dependent samples

Code	N	d	SED	t	q′	S′
WXY.1, WXY.2	102.3	.155	.359	0.43		0.37
WXY.1, WXY.3	102.3	.138	.354	0.39		0.33
WXY.2, WXY.3	102.3	.017	.363	0.05		0.04

Table B.14.2 Comparison of position-group interactions

Code	N	d	SED	t	q′	S′
W1.1, X1.3	103.5	.260	.354	0.73	0.45	
W1.1, Y1.2	101	.090	.346	0.26	0.16	
X1.3, Y1.2	102.5	.350	.360	0.97	0.61	0.49
W2.2, X2.1	103.5	.671	.380	1.76	1.17	0.93
W2.2, Y2.3	101	.270	.362	0.75	0.47	
X2.1, Y2.3	102.5	.401	.373	1.07	0.70	0.56
W3.3, X3.2	103.5	.570	.368	1.55	1.00	0.79
W3.3, Y3.1	101	.270	.332	0.81	0.47	0.37
X3.2, Y3.1	102.5	.300	.349	0.86	0.52	0.42

Table B.15.1 Comparison of forms for dependent samples

Code	N	d	SED	t	q′	S′
Z1, 2	307	.405	.127	3.20		0.96
Z1, 3	307	.045	.132	0.34		0.14
Z2, 3	307	.450	.126	3.58		1.10

Table B.15.2 Comparison of form-position interactions

Code	N	d	SED	t	q′	Q′
W1.1, 2.2	102	.140	.214	0.65		0.39
X2.1, 1.3	105	.551	.224	2.46	0.96	1.53
Y1.2, 2.3	100	.420	.227	2.20	0.87	1.17
W1.1, 3.3	102	.281	.223	1.26		0.78
X3.2, 1.3	105	.030	.231	0.13		0.08
Y3.1, 1.2	100	.080	.235	0.34		0.22
W2.2, 3.3	102	.420	.225	1.87		1.17
X2.1, 3.2	105	.521	.206	2.52	1.04	1.45
Y3.1, 2.3	100	.500	.230	1.83		1.39

Table B.16.1 Comparison of groups for independent samples

Code	N	d	SED	t	q′	S′
W123, X123	103.5	1.500	.961	1.56		1.20
W123, Y123	101	.450	.883	0.51		0.36
X123, Y123	102.5	1.049	.931	1.13		0.84

Table B.16.2 Comparison of group-form interactions

Code	N	d	SED	t	q′	S′
W3.3, X1.3	103.5	.540	.360	1.50		
W3.3, Y2.3	101	.690	.354	1.95		
X1.3, Y2.3	102.5	.150	.361	0.42		
W2.2, X3.2	103.5	.150	.375	0.40		
W2.2, Y1.2	101	.230	.361	0.64		
X3.2, Y1.2	102.5	.380	.368	1.03		
W1.1, X2.1	103.5	.810	.367	2.21	1.41	1.12
W1.1, Y3.1	101	.011	.326	0.03	0.02	0.02
X2.1, Y3.1	102.5	.821	.354	2.32	1.43	1.14

Table B.17.1 Comparison of variances by group for independent samples

Code	df	s^2	s^2	F	P less than
W123, X123	100, 103	42.7	52.8	1.24	.15
W123, Y123	100, 98	42.7	36.0	1.19	.20
X123, Y123	103, 98	52.8	36.0	1.47	.03

Table B.17.2 Comparison of variances for dependent samples

Code	df	r	Δs^2	t	P less than
W1.1, 2.2	100	.6364	1.02	1.31	
X2.1, 1.3	103	.6510	0.93	1.06	
Y1.2, 2.3	98	.5890	0.09	0.11	
Z1, Z2	305	.6273	0.50	1.09	
W1.1, 3.3	100	.5871	0.44	0.55	
X3.2, 1.3	103	.6186	0.59	0.66	
Y3.1, 1.2	98	.5033	1.39	1.81	.08
Z1, Z3	305	.5731	0.10	0.21	
W2.2, 3.3	100	.6133	0.58	0.70	
X2.1, 3.2	103	.7145	0.34	0.41	
Y3.1, 2.3	98	.5278	1.48	1.96	.06
Z2, Z3	305	.6299	0.60	1.32	

Table B.18 Standard errors and critical values for statistical tests of the significance of differences in means

Table	I*	SED - Critical S		SED - Critical q		SED - Critical Q	
B.4	D			.271	5.30	.153	5.30
B.5	D			.271	5.30	.153	5.30
B.6	I	.271	6.27				
B.7	D	.346	3.75	.245	3.85		
B.8	I	.383	6.27	.271	5.30		
B.9	I	.383	6.27	.271	5.30		
B.14.1	D	.207	4.00				
B.14.2	I	.359	4.00	.254	4.49		
B.15.1	D	.207	4.00				
B.15.2	D			.254	4.49	.160	4.49
B.16.1	I	.207	4.00				
B.16.2	I	.359	4.00	.254	4.49		

*Independent uncorrelated samples are indicated 'I' and dependent or correlated samples are indicated 'D'

Appendix C Miscellaneous tables

C.1. Summary data for the calculation of the Stability Estimate of the Mean

The data presented in this table are the basis for the results listed in Table 2.9 on p. 136 of Section II.D.4. These statistics are derived from Table B.3 in the previous appendix. The split differences in the means, $^{d}/_{2}$, are given in units of percent.

C.2. Composite reliability of a series of tests

The reliability of a composite of several tests is calculated from the intercorrelations among the tests and their average reliability by the equation

$$r_n = \frac{\bar{r} + (n-1)\bar{r}_{ij}}{1 + (n-1)\bar{r}_{ij}}$$

where n is the number of tests combined to give the overall score, $\bar{r}$ is the mean reliability of these tests, and $\bar{r}_{ij}$ is the mean test intercorrelation. It is assumed that the test reliabilities and intercorrelations are not widely scattered but are closely clustered.

When tests are uncorrelated, the cumulative reliability is not higher than the mean test reliability. When test intercorrelations equal the reliabilities (they cannot logically be larger), this formula simplifies to the Spearman-Brown prophecy formula. This formula may also be used to calculate the reliability of a test of n items based on the item reliabilities and item intercorrelations when some method other than item intercorrelation is used to derive item reliability.

C.3. Intercorrelations of different unit exams and standard deviation of term average

The values in this table include the intercorrelations of unit tests and the standard deviation of the last-tries mean in units of percent. The intercorrelations for all tries is provided because the last-tries intercorrelations are artificially higher since students retake exams to raise their test scores to similar levels. Yet the all-tries intercorrelation is an underestimate because many students take an exam before they have reached their final level of competence. The 'true' intercorrelations of the unit tests lies between these two extremes.

The values presented in this table may be used in conjunction with Table C.2 to calculate the cumulative reliability of various term averages for CEM 130 and 131.

Table C.1 Summary data for the calculation of the Stability Estimate of the Mean

Code	N	ICS	$\overline{KR}_{20}$	$d/2$
A1	296	.213	.6383	2.120
A2	284	.212	.6079	0.315
A3	277	.212	.6682	1.455
D1	72	.175	.6689	0.650
D2	76	.175	.6928	5.040
E1	86	.185	.7330	4.570
E2	62	.190	.7626	0.270
E3	73	.182	.6914	1.600
E4	58	.190	.7351	4.370
J1	109	.147	.6248	0.395
J2	112	.147	.6795	0.325
L1	86	.158	.6853	2.250
L2	80	.158	.6276	2.000
N1	170	.112	.6080	1.270
N3	153	.111	.6178	4.160
K1	66	.037	.7512	2.525
K2	65	.037	.7964	1.740
H1	40	.026	.7599	4.500
H2	44	.026	.7766	2.955

Table C.2 Composite reliability of a series of tests

$\bar{r}$	t	$\bar{r}_{ij}$						
		.1500	.2000	.2500	.3000	.3500	.4000	$\bar{r}_{ij} = \bar{r}$
.6000	2	.6522	.6667	.6800	.6923	.7037	.7143	.7500
.6000	3	.6923	.7143	.7333	.7500	.7647	.7778	.8182
.6000	4	.7241	.7500	.7714	.7895	.8049	.8182	.8571
.6000	5	.7500	.7778	.8000	.8182	.8333	.8462	.8824
.6000	6	.7714	.8000	.8222	.8400	.8545	.8667	.9000
.6000	7	.7895	.8182	.8400	.8571	.8710	.8824	.9138
.6000	8	.8049	.8333	.8545	.8710	.8841	.8947	.9231
.6000	9	.8182	.8462	.8667	.8824	.8947	.9048	.9310
.6000	10	.8298	.8571	.8769	.8919	.9036	.9130	.9375
.6000	11	.8400	.8667	.8857	.9000	.9111	.9200	.9429
.6900	2	.7304	.7417	.7520	.7615	.7704	.7786	.8166
.6900	3	.7615	.7786	.7933	.8063	.8176	.8278	.8697
.6900	4	.7862	.8063	.8229	.8368	.8488	.8591	.8990
.6900	5	.8063	.8278	.8450	.8591	.8708	.8808	.9176
.6900	6	.8229	.8450	.8622	.8760	.8873	.8967	.9303
.6900	7	.8368	.8591	.8760	.8893	.9000	.9088	.9397
.6900	8	.8488	.8708	.8873	.9000	.9101	.9184	.9468
.6900	9	.8591	.8808	.8967	.9088	.9184	.9262	.9525
.6900	10	.8681	.8893	.9046	.9162	.9253	.9326	.9570
.6900	11	.8760	.8967	.9114	.9225	.9311	.9380	.9608
.7500	2	.7826	.7917	.8000	.8077	.8148	.8214	.8571
.7500	3	.8077	.8214	.8333	.8438	.8529	.8611	.9000
.7500	4	.8276	.8438	.8571	.8684	.8780	.8864	.9231
.7500	5	.8438	.8611	.8750	.8864	.8958	.9038	.9375
.7500	6	.8571	.8750	.8889	.9000	.9091	.9167	.9474
.7500	7	.8684	.8864	.9000	.9107	.9194	.9265	.9545
.7500	8	.8780	.8958	.9091	.9194	.9275	.9342	.9600
.7500	9	.8864	.9038	.9167	.9265	.9342	.9405	.9643
.7500	10	.8936	.9107	.9231	.9324	.9398	.9457	.9677
.7500	11	.9000	.9167	.9286	.9375	.9444	.9500	.9706
.7900	2	.8174	.8250	.8320	.8385	.8444	.8500	.8827
.7900	3	.8385	.8500	.8600	.8688	.8765	.8833	.9186
.7900	4	.8552	.8688	.8800	.8895	.8976	.9045	.9377
.7900	5	.8688	.8833	.8950	.9045	.9125	.9192	.9495
.7900	6	.8800	.8950	.9067	.9160	.9236	.9300	.9576
.7900	7	.8895	.9045	.9160	.9250	.9323	.9382	.9674
.7900	8	.8976	.9125	.9236	.9323	.9391	.9447	.9678
.7900	9	.9045	.9192	.9300	.9382	.9447	.9500	.9713
.7900	10	.9106	.9250	.9354	.9432	.9494	.9543	.9741
.7900	11	.9160	.9300	.9400	.9475	.9533	.9580	.9764
.8100	2	.8348	.8417	.8480	.8538	.8593	.8643	.8950
.8100	3	.8538	.8643	.8733	.8813	.8882	.8944	.9275
.8100	4	.8690	.8813	.8914	.9000	.9073	.9136	.9446
.8100	5	.8813	.8944	.9050	.9136	.9208	.9269	.9552
.8100	6	.8914	.9050	.9156	.9240	.9309	.9367	.9624
.8100	7	.9000	.9136	.9240	.9321	.9387	.9441	.9676
.8100	8	.9073	.9208	.9309	.9387	.9449	.9500	.9715
.8100	9	.9136	.9269	.9367	.9441	.9500	.9548	.9746
.8100	10	.9191	.9321	.9415	.9486	.9542	.9587	.9771
.8100	11	.9240	.9367	.9457	.9525	.9578	.9620	.9791

Table C.3 Intercorrleations of different unit exams and standard deviation of term average

Term	No. of Exams	Tries per Exam	Intercorrelations All Tries	Intercorrelations Last Tries	Term average σ
130W73	8	1.744	.30	.46	10.129
131S73	7	2.076	.28	.46	11.823
130F73	9	1.580	.27	.43	9.708
131W74	5	1.993	.32	.50	12.250
130W74	8	1.795	.37	.53	11.894
131S74	5	1.974	.37	.55	13.722
130F74	8	1.785	.33	.50	9.618
131W75	5	2.201	.34	.51	11.861
130W75	6	2.132	.35	.52	11.579
131S75	5	2.242	.32	.45	12.131
130F75	6	2.103	.35	.55	10.850
131W76	5	2.694	.29	.54	13.656
130W76	6	2.019	.36	.52	12.472
131S76	5	2.097	.35	.47	12.584
mean			.33	.50	11.734

Appendix D Statistics for subsets of items from the fifteen-item tests of various comparison schemes

For none of the comparison schemes were all fifteen test questions of identical format. To investigate the various effects of item format and content on test reliability, statistics are given for subtests of items more nearly alike in the number of answer choices or the types of questions. Only two categories of 'content' are used to classify questions -- problems and nonproblems. The rationale for this distinction is explained in Section II.D.

Table D.1 Test statistics* adjusted to a standard ten-item length

Code	FP	ICS	$_{10}s^2$	$_{10}r$	$_{10}r'$
A1	.60	.213	3.603	.5042	.7555
A2	.60	.212	3.473	.4556	.7126
A3	.60	.212	3.762	.5593	.7993
D1	.80	.175	4.008	.6385	.8454
D2	.80	.175	4.195	.6728	.8668
E1	.87	.185	4.443	.5534	.6758
E2	.73	.190	4.757	.4799	.6900
E3	.73	.182	4.326	.4826	.6984
E4	.73	.190	4.695	.4853	.6960
L1	.00	.158	4.365	.6518	.8321
L2	.00	.158	3.657	.4530	.6605
J1	.20	.147	3.090	.4804	.6636
J2	.20	.147	3.834	.6926	.8513
N1	.57	.112	3.525	.5069	.6557
N2	.60	.111	3.509	.5295	.6665
G1	.47	.084	5.316	.5925	.7022
G2	.47	.084	4.867	.7020	.8157
K1	.73	.037	5.283	.6796	.7216
K2	.73	.037	6.275	.8124	.8478
H1	.80	.026	5.636	.7780	.8023
H2	.80	.026	5.308	.8061	.8291

*Taken from Table B.3

Table D.2 Statistics for subtests of problems

Code	k	ICS	$_{10}s^2$	$_{10}r$	$_{10}r'$
A1	7	.200	4.034	.6686	.8598
A2	7	.200	4.105	.6551	.8382
A3	7	.200	4.395	.6920	.8713
D1	10	.200	4.025	.6277	.8451
D2	10	.200	3.900	.7301	.9231
E1	9	.200	4.340	.5453	.7696
E2	9	.200	3.555	.4920	.7378
E3	9	.200	4.648	.5373	.7608
E4	9	.200	4.916	.5497	.7675
N1	8	.100	4.251	.6792	.7897
N2	8	.100	4.301	.6830	.7834
G1	5	.050	5.011	.8071	.8575
G2	5	.050	5.311	.7852	.8397
K1	11	.000	6.167	.7569	.7569
K2	11	.000	7.355	.8418	.8418
H1	12	.000	6.924	.8146	.8146
H2	12	.000	5.600	.8100	.8100

Table D.3 Statistics for subtests of nonnumerical questions

Code	k	ICS	$_{10}s^2$	$_{10}r$	$_{10}r'$
A1	6	.233	2.597	.4055	.7535
A2	6	.222	2.561	.4632	.7802
A3	6	.222	2.269	.4407	.7942
B	4	.233	4.031	.6204	.8695
C1	4	.244	1.971	.5461	.8716
C2	4	.227	2.945	.5000	.8431
E1	4	.200	3.352	.5068	.7874
E2	4	.200	4.373	.4723	.7344
E3	4	.200	3.116	.5350	.8342
E4	4	.200	2.986	.6302	.8958
L1	7	.175	3.736	.6075	.8180
L2	7	.175	2.617	.2377	.5236
J1	9	.167	2.637	.5588	.7671
J2	9	.167	2.874	.5965	.8011
L1'	8	.141	4.412	.7878	.9432
L2'	8	.141	5.127	.6878	.8446
K1	4	.137	3.126	.7282	.8886
K2	4	.137	4.074	.7409	.8761

Table D.3 (cont'd)

Code	k	ICS	$_{10}s^2$	$_{10}r$	$_{10}r'$
H1	3	.132	3.668	.7433	.9251
H2	3	.132	3.787	.8486	.9964
I	10	.130	1.639	.5125	.7822
N1	7	.125	2.798	.4190	.6280
N2	7	.125	2.842	.4745	.6660
F	6	.103	3.028	.4540	.6593
I'	5	.000	4.551	.6772	.6772

Table D.4 Statistics for subtests with identical content

Code	k	ICS	$_{10}s^2$	$_{10}r$	$_{10}r'$
B	6	.100	5.521	.7914	.8649
C1	6	.100	3.450	.6543	.7774
C2	7	.100	4.250	.7354	.8742
F	6	.050	6.553	.8847	.9219
B'	5	.100	2.960	.6372	.7826
C1'	5	.119	3.583	.7255	.8623
C2'	4	.119	2.830	.6238	.7722

Table D.5.1 Variance fractions for fifteen-item tests

Code	ICS	VF_T	VF_e	VF_g	VF_R
A, D, E	.193	.6230	.3770	.1930	.1840
L, J, N	.139	.6465	.3535	.1469	.2067
G	.084	.7325	.2675	.0919	.1757
K, H	.032	.8324	.1676	.0243	.1433

Table D.5.2 Partial variances for fifteen-item tests

Code	s_t^2	s_T^2	s_e^2	s_g^2	s_R^2
A, D, E	8.080	5.024	3.057	1.548	1.509
L, J, N	6.996	4.571	2.425	1.023	1.402
G	10.077	7.356	2.722	0.927	1.795
K, H	11.363	9.480	1.883	0.276	1.607

Table D.5.3 Selected statistics* with guessing partialled out

Code	$VF_{T'}$	$VF_{R'}$	$s_{t'}^2$	$s_{T'}^2$	$s_{R'}^2$
A, D, E	.8101	.1899	12.025	9.742	2.283
L, J, N	.8097	.1903	10.054	8.141	1.914
G	.8360	.1640	13.771	11.513	2.259
K, H	.8722	.1278	13.651	11.906	1.745

*From Tables D.5.1 and D.5.2

Table D.6.1 Variance fractions for subtests of problems

Code	ICS	VF_T	VF_e	VF_g	VF_R
A, D, E	.200	.6109	.3891	.2104	.1787
N	.100	.6811	.3189	.1055	.2134
G	.050	.7961	.2039	.0525	.1514
H, K	.000	.8058	.1942	.0000	.1942

Table D.6.2 Partial variances for subtests of problems

Code	s_t^2	s_T^2	s_e^2	s_g^2	s_R^2
A, D, E	4.213	2.568	1.645	0.886	0.759
N	4.276	2.912	1.364	0.451	0.913
G	5.161	4.107	1.054	0.271	0.783
H, K	6.512	5.259	1.253	0.000	1.253

Table D.6.3 Selected statistics* with guessing partialled out

Code	$VF_{T'}$	$VF_{R'}$	$s_{t'}^2$	$s_{T'}^2$	$s_{R'}^2$
A, D, E	.8106	.1894	6.346	5.144	1.202
N	.8069	.1931	5.897	4.758	1.139
G	.8619	.1381	6.653	5.734	0.919
H, K	.8058	.1942	6.512	5.529	1.253

*From Tables D.6.1 and D.6.2

Table D.7.1 Variance Fractions for subtests of nonproblems

Code	ICS	VF_T	VF_e	VF_g	VF_R
A	.226	.4365	.5635	.3395	.2240
L, J	.171	.5001	.4999	.2273	.2726
L', I, N, F	.128	.5559	.4441	.1980	.2461
I'	.000	.6772	.3228	.0000	.3228

Table D.7.2 Partial variances for subtests of nonproblems

Code	s_t^2	s_T^2	s_e^2	s_g^2	s_R^2
A	2.476	1.080	1.396	0.849	0.547
L, J	2.966	1.521	1.450	0.668	0.782
L', I, N, F	3.307	1.956	1.352	0.609	0.743
I'	4.551	3.082	1.469	0.000	1.469

Table D.7.3 Selected statistics* with guessing partialled out

Code	$VF_{T'}$	$VF_{R'}$	$s_{t'}^2$	$s_{T'}^2$	$s_{R'}^2$
A	.7464	.2536	3.830	2.859	0.972
L, J	.7462	.2538	4.397	3.281	1.116
L', I, N, F	.7841	.2159	4.852	3.804	1.048
I'	.6772	.3228	4.551	3.082	1.469

*From Tables D.7.1 and D.7.2

Table D.8.1 Selected statistics for subtests with identical content

Code	ICS	s^2	VF_T	VF_g	VF_R
B, C1, C2	.100	4.407	.7330	.1058	.1612
F	.050	6.553	.8847	.0372	.0781
B', C1', C2'	.113	3.124	.6622	.1435	.1943

Table D.8.2 Selected statistics* with guessing partialled out

Code	s^2	s^2	s^2	$VF_{T'}$	$VF_{R'}$
B, C1, C2	0.683	0.914	6.130	.8509	.1491
F	0.472	0.614	8.368	.9266	.0734
B', C1', C2'	0.594	0.818	4.490	.8178	.1822

*From Table D.8.1

Table D.9.1 Linear equations for subtests with identical content

Number of points used	Dependent variable	Equation	Coefficient of determination
7	s_t^2	8.9481 − 48.677(ICS)	.65
7	s_T^2	8.3267 − 52.925(ICS)	.68
7	s_e^2	0.6214 + 4.2473(ICS)	.48
7	s_g^2	0.1075 + 3.1622(ICS)	.71
7	s_R^2	uncorrelated	.02
7	VF_e	-0.0285 + 3.0944(ICS)	.57
7	VF_g	-0.0384 + 1.5321(ICS)	.73

Table D.9.2 Linear equations for fifteen-item tests

Number of points used	Dependent variable	Equation	Coefficient of determination
21 (excl. I)	s_t^2	11.285 − 19.814(ICS)	.47
21	s_T^2	9.5201 − 26.092(ICS)	.63
21	s_e^2	1.7652 + 6.2786(ICS)	.28
21	s_g^2	0.0851 + 7.4246(ICS)	.88
21	s_R^2	uncorrelated	.02
21 (excl. I)	s_t^2	6.3222 + 3.753(FP)	.25
21	s_T^2	4.0268 + 3.179(FP)	.16
21	FP	0.642 − 0.506(ICS)	.02
22	VF_e	0.1717 + 1.1390(ICS)	.47
21 (excl. I)	VF_e	0.1528 + 1.2272(ICS)	.55
4 means*	VF_e	0.1445 + 1.3205(ICS)	.92
21 (excl. I)	VF_g	-0.0025 + 1.0313(ICS)	.95

*The four groups are: 1) A, D, E; 2) L, J, N; 3) G; 4) K, H

Table D.9.3 Linear equations for subtests of problems

Number of points used	Dependent variable	Equation	Coefficient of determination
17	s_t^2	6.1302 − 10.176(ICS)	.69
17	s_T^2	4.9688 − 12.515(ICS)	.80
17	s_e^2	1.1613 + 2.3397(ICS)	.30
17	s_g^2	0.0142 + 4.3704(ICS)	.95
17	s_R^2	1.1394 − 1.7618(ICS)	.20
17	VF_e	0.1884 + 1.0103(ICS)	.67
17	VF_g	0.0000 + 1.0520(ICS)	.98

Table D.9.4 Linear equations for subtests of nonproblems

Number of points used	Dependent variable	Equation	Coefficient of determination
4 means	VF_g	-0.0003 + 1.4592(ICS)	.99
24 (excl. I')	VF_g	0.0122 + 1.3132(ICS)	.67
24 (excl. I')	VF_e	0.2446 + 1.1067(ICS)	.11
25	VF_e	0.2782 + 0.9237(ICS)	.13
24 (excl. I')	s_g^2	0.1939 + 3.1054(ICS)	.51
24 (excl. I')	s_e^2	uncorrelated	.05
25	s_e^2	uncorrelated	.02

Appendix E Testing and grading statistics for CEM 130 and 131

Statistics on the test means, test taking frequencies, and mean grades before and after the final exam for enrollments of 300 or more students are presented in this appendix. All data are given as percent unless specified.

Certain students are excluded from consideration in these tables, based on the number of exams they missed. Some students unofficially drop a course by no longer attending class or taking tests. The scores of zero entered in these students' records do not represent a failure to learn after an attempt was made but simply a failure to attempt to learn. This author decided that to include such scores in calculated exam means and grade averages would distort the results, and so such students were deleted by the computer programs which processed these data. The fact that these students almost uniformly receive a grade of 0.0 from the registrar means that the average grade based on the total number of grades assigned by the university will be lower than the average grade given here. This difference is typically about 0.04; almost half the differences are 0.01 or less.

Table E.1 Testing and grading statistics for CEM130W73

Exam	Total mean	Net mean	No. of tries	Distribution of exam-taking frequency					
				1	2	3	4	5	6
1	74.56	79.57	1.52	61	30	7	1	0	0
2	67.08	76.30	1.71	50	34	13	3	0	0
3	75.48	81.70	1.58	59	27	11	2	0	0
4	76.27	84.74	1.57	61	26	8	4	1	0
5	72.39	79.63	1.79	46	36	12	4	1	0
6	77.18	84.56	1.72	51	32	11	4	1	0
7	63.59	74.64	2.27	32	29	22	10	4	1
8	64.31	73.76	1.81	48	30	14	3	2	1
all	71.81	79.28	1.74	51	31	12	4	1	0

N	Final exam mean	G. P. A. before final	G. P. A. after final
556	66.82	3.08	2.85

	Number	Percent
Grades lowered by final exam:	259	46.58
Grades unchanged by final exam:	291	52.34
Grades raised by final exam:	6	1.08

Table E.2 Testing and grading statistics for 131S73

Exam	Total mean	Net mean	No. of tries	Distribution of exam-taking frequency 1	2	3	4	5	6
1	67.58	77.08	2.33	27	33	24	11	4	1
2	66.69	76.28	2.28	34	26	21	11	5	2
3	68.38	76.87	2.20	35	30	21	9	3	2
4	72.31	79.75	1.77	50	31	12	5	1	0
5	66.37	75.03	2.03	42	28	19	7	4	0
6	64.66	73.91	1.97	42	31	17	6	3	1
7	52.44	61.20	1.95	43	29	17	7	1	1
all	65.93	74.37	2.08	39	30	19	8	3	1

N	Final exam mean	G. P. A. before final	G. P. A. after final
498	40.65	2.72	2.53
		Number	Percent.
Grades lowered by final exam:		178	35.74
Grades unchanged by final exam:		319	64.06
Grades raised by final exam:		1	0.20

Table E.3 Testing and grading statistics for CEM130F73

Exam	Total mean	Net mean	No. of tries	Distribution of exam-taking frequency 1	2	3	4	5	6
1	78.02	83.54	1.55	58	31	9	2	0	0
2	75.74	82.38	1.56	56	32	10	1	0	0
3	80.80	85.90	1.48	61	30	7	1	0	0
4	81.68	88.09	1.50	63	27	7	2	0	0
5	76.32	80.92	1.59	58	28	10	3	0	0
6	79.82	84.68	1.50	60	32	6	2	0	0
7	68.36	74.65	1.83	47	31	15	5	1	0
8	74.39	82.09	1.62	53	33	10	2	0	0
9	68.90	75.86	1.61	58	27	10	3	1	0
all	76.36	82.04	1.58	57	30	9	2	0	0

N	Final exam mean	G. P. A. before final	G. P. A. after final
1221	74.66	3.27	3.17

	Number	Percent
Grades lowered by final exam:	307	25.14
Grades unchanged by final exam:	855	70.02
Grades raised by final exam:	60	4.91

Table E.4 Testing and grading statistics for CEM131W74

Exam	Total mean	Net mean	No. of tries	Distribution of exam-taking frequency 1	2	3	4	5	6
1	74.04	80.16	1.84	44	35	16	5	1	0
2	71.89	80.24	2.02	38	31	22	6	2	0
3	71.16	77.11	1.89	44	31	18	5	1	0
4	68.51	77.02	1.99	36	35	22	5	1	0
5	57.98	66.10	2.24	28	34	26	9	3	0
all	68.72	76.31	1.99	38	33	21	6	1	0

N	Final exam mean	G. P. A. before final	G. P. A. after final
1110	63.46	2.85	2.64
		Number	Percent
Grades lowered by final exam:		503	45.32
Grades unchanged by final exam:		571	51.44
Grades raised by final exam:		36	3.24

Table E.5 Testing and grading statistics for CEM130W74

Exam	Total mean	Net mean	No. of tries	Distribution of exam-taking frequency 1	2	3	4	5	6
1	71.12	78.33	1.72	48	36	11	3	1	0
2	73.79	78.52	1.52	61	28	9	2	0	0
3	73.80	79.84	1.65	55	28	13	3	1	0
4	69.92	74.82	1.98	42	31	15	9	2	1
5	71.98	76.52	1.55	62	25	10	3	1	0
6	63.74	70.63	2.02	38	33	18	7	2	1
7	75.13	79.49	1.82	50	27	14	6	2	0
8	59.49	64.13	2.12	38	28	20	8	3	1
all	69.61	75.42	1.80	49	30	14	5	1	0

N	Final exam mean	G. P. A. before final	G. P. A. after final
780	73.37	2.79	2.80

	Number	Percent
Grades lowered by final exam:	82	10.51
Grades unchanged by final exam:	595	76.28
Grades raised by final exam:	103	13.21

Table E.6 Testing and grading statistics for CEM131S74

Exam	Total mean	Net mean	No. of tries	Distribution of exam-taking frequency 1	2	3	4	5	6
1	74.01	81.18	1.99	40	32	18	7	2	0
2	73.19	80.67	1.83	44	34	16	4	1	0
3	65.30	72.53	1.98	42	29	20	7	2	0
4	64.95	73.67	1.94	38	38	16	7	1	0
5	55.30	63.85	2.13	32	35	21	9	2	0
all	66.79	74.45	1.97	39	34	19	7	2	0

N	Final exam mean	G. P. A. before final	G. P. A. after final
660	61.81	2.72	2.51

	Number	Percent
Grades lowered by final exam:	289	43.79
Grades unchanged by final exam:	350	53.03
Grades raised by final exam:	21	3.18

Table E.7 Testing and grading statistics for CEM130F74

Exam	Total mean	Net mean	No. of tries	Distribution of exam-taking frequency 1	2	3	4	5	6
1	71.11	76.60	1.73	48	34	13	4	1	0
2	75.51	81.55	1.57	57	30	10	2	0	0
3	74.77	81.25	1.67	50	35	12	3	0	0
4	75.91	81.73	1.74	51	29	14	5	1	0
5	73.80	80.73	1.72	49	34	13	4	1	0
6	72.40	79.71	1.74	47	35	14	3	0	0
7	76.24	82.80	2.00	37	37	16	7	3	0
8	70.30	78.18	2.10	40	30	17	7	3	2
all	74.18	80.45	1.79	48	33	14	4	1	0

N	Final exam mean	G. P. A. before final	G. P. A. after final
1242	58.63	2.87	2.65

	Number	Percent
Grades lowered by final exam:	557	44.85
Grades unchanged by final exam:	656	52.82
Grades raised by final exam:	29	2.33

Table E.8 Testing and grading statistics for CEM131W75

Exam	Total mean	Net mean	No. of tries	Distribution of exam-taking frequency 1	2	3	4	5	6
1	72.73	79.70	1.87	43	33	18	5	1	0
2	73.09	82.82	2.16	35	34	16	9	4	1
3	65.00	73.76	2.37	27	33	23	12	4	2
4	66.31	75.63	2.27	27	34	26	9	3	1
5	60.43	69.59	2.34	27	32	25	11	3	1
all	67.87	76.64	2.20	32	33	22	9	3	1

N	Final exam mean	G. P. A. before final	G. P. A. after final
1112	66.45	2.64	2.50

	Number	Percent
Grades lowered by final exam:	346	31.12
Grades unchanged by final exam:	703	63.22
Grades raised by final exam:	63	5.67

Table E.9 Testing and grading statistics for CEM130W75

Exam	Total mean	Net mean	No. of tries	Distribution of exam-taking frequency 1	2	3	4	5	6
1	64.69	73.23	2.06	33	35	22	7	2	0
2	72.49	82.74	1.89	43	35	14	6	1	1
3	71.33	81.53	2.00	37	36	18	7	1	0
4	67.75	76.39	2.26	32	27	25	13	3	0
5	66.25	75.01	2.20	33	32	21	9	4	1
6	60.11	70.71	2.38	27	30	27	11	4	1
all	67.81	76.70	2.13	34	33	21	9	3	1

N	Final exam mean	G. P. A. before final	G. P. A. after final
634	71.42	2.66	2.62
		Number	Percent
Grades lowered by final exam:		107	16.88
Grades unchanged by final exam:		463	73.03
Grades raised by final exam:		64	10.09

Table E.10 Testing and grading statistics for CEM131S75

Exam	Total mean	Net mean	No. of tries	Distribution of exam-taking frequency 1	2	3	4	5	6
1	67.17	78.88	2.33	28	30	24	13	3	1
2	69.98	80.21	2.09	35	30	24	9	1	0
3	61.97	74.22	2.27	31	28	27	10	3	1
4	59.99	70.90	2.19	35	29	21	10	3	1
5	50.94	60.02	2.33	31	30	21	10	5	2
all	63.45	73.32	2.24	32	29	23	10	3	1

N	Final exam mean	G. P. A. before final	G. P. A. after final
545	59.50	2.37	2.18

	Number	Percent
Grades lowered by final exam:	232	42.57
Grades unchanged by final exam:	283	51.93
Grades raised by final exam:	30	5.50

Table E.11 Testing and grading statistics for CEM130F75

Exam	Total mean	Net mean	No. of tries	Distribution of exam-taking frequency 1	2	3	4	5	6
1	68.49	79.06	2.08	34	34	23	8	1	0
2	75.88	84.62	1.84	44	36	13	5	1	0
3	73.59	82.58	2.01	39	32	20	8	1	1
4	74.04	82.22	2.15	31	35	22	10	2	0
5	71.42	81.89	2.07	38	31	19	9	2	1
6	64.40	74.98	2.47	25	30	24	13	5	2
all	71.99	80.91	2.10	35	33	20	9	2	1

N	Final exam mean	G. P. A. before final	G. P. A. after final
1381	70.34	2.99	2.83
		Number	Percent
Grades lowered by final exam:		476	34.47
Grades unchanged by final exam:		855	61.91
Grades raised by final exam:		50	3.62

Table E.12 Testing and grading statistics for CEM131W76

Exam	Total mean	Net mean	No. of tries	Distribution of exam-taking frequency 1	2	3	4	5	6
1	71.79	82.83	2.31	29	31	24	11	4	1
2	72.57	81.60	2.12	36	31	22	9	2	0
3	62.49	72.43	2.46	25	30	25	14	5	1
4	61.25	72.29	2.48	23	32	25	13	5	1
5	57.12	75.34	4.10	2	7	24	26	23	18
all	64.04	77.11	2.69	23	26	24	15	8	4

N	Final exam mean	G. P. A. before final	G. P. A. after final
1267	66.42	2.70	2.54
		Number	Percent
Grades lowered by final exam:		453	35.75
Grades unchanged by final exam:		770	60.77
Grades raised by final exam:		44	3.47

Table E.13 Testing and grading statistics for CEM130W76

Exam	Total mean	Net mean	No. of tries	Distribution of exam-taking frequency 1	2	3	4	5	6
1	66.98	74.54	2.09	34	34	21	9	2	0
2	72.49	80.09	1.83	46	33	14	5	1	0
3	72.00	81.31	1.97	41	33	17	6	2	1
4	74.25	80.42	1.81	49	28	15	7	1	0
5	68.46	75.64	1.96	39	36	16	6	2	0
6	57.60	65.33	2.46	25	32	23	12	5	2
all	69.18	76.79	2.02	39	33	18	8	2	1

N	Final exam mean	G. P. A. before final	G. P. A. after final
671	75.04	2.65	2.69
		Number	Percent
Grades lowered by final exam:		67	9.99
Grades unchanged by final exam:		476	70.94
Grades raised by final exam:		128	19.08

Table E.14 Testing and grading statistics for CEM131S76

Exam	Total mean	Net mean	No. of tries	Distribution of exam-taking frequency 1	2	3	4	5	6
1	69.05	78.60	2.11	33	34	22	8	2	0
2	71.39	79.54	1.94	41	32	19	7	1	0
3	60.54	70.51	2.31	27	35	22	10	4	2
4	60.92	72.64	2.05	37	33	19	8	2	0
5	51.93	60.65	2.07	42	28	18	6	3	2
all	64.24	72.98	2.10	36	32	20	8	3	1

N	Final exam mean	G. P. A. before final	G. P. A. after final
575	67.14	2.39	2.35

	Number	Percent
Grades lowered by final exam:	108	18.78
Grades unchanged by final exam:	400	69.57
Grades raised by final exam:	67	11.65

Appendix F Attitude survey and course evaluation results

Four tables are presented in this appendix. The first two tables are based on the total response to course evaluation surveys; the statistics were calculated by hand. The last two tables are based only on those surveys matched to records in the course files by student number. Each table is briefly described below.

F.1. Fourteen graphs of average response to Items 1 through 14 from the course evaluation form indicating mean and interpolated median responses for each of fourteen terms with major enrollments

In this table are listed graphically for each item the mean and interpolated median responses for terms with enrollments of 300 or more students. The interpolated median is based on the assumption that responses within each interval would be evenly distributed among subintervals if they were available. When the mean and median are different, then the distribution of the responses is skewed and the response mean may not be the best estimate of the 'average' response.

F.2. Twenty-four bar graphs of percent response in each response category for Items 1 through 14 and 18 through 22 for terms with major enrollments

In this table are listed the percent response in each response category to items from the course evaluation survey for terms with enrollments of 300 or more students. The bar graphs for all but the last item are presented in the following format.

The identifying numbers and ordering of the response percent blocks are listed to the left of each bar graph. Bar widths are proportionsl to the number of responses to each survey question. Intervals of ten percent are marked on the edges of the graphs and the midpoint in the distribution is indicated by an underlying dotted line. Each survey question is provided underneath the appropriate graph; additional reference may be made to Figure 4.4 on p. 214 for a sample of a typical survey. Since Items 18 to 22 are not always the same for different terms, those questions which are different from the ones which appear on the sample survey are described here.

Question 19 on p. 372 asks the student which textbook was of significant value. Four different texts have been used over the past five years, and for the terms this question was asked, these books have the following appearances:

Mortimer - Chemistry A Conceptual Approach 

Petrucci - General Chemistry 

O'Connor - Fundamentals of Chemistry 

Brady, Humiston - General Chemistry Principles and Structure 

Question 18 on p. 374 pertains to the pacing of presentation:
The pacing in the course should be: (1) faster (2) about the same or (3) slower.

Question 19 on p. 374 asks about examination length:
The exams should be: (1) shorter (2) about as they are now or (3) longer and less often.

Question 20 on p. 374 asks the student whether:
The number of Study Questions should be: (1) increased (2) unchanged or (3) decreased.

Question 21 on p. 374 pertains to the number of course credits: This course should be: (1) lowered to 3 credits (2) remain 4 credits or (3) be increased to 5 credits. (When asked of CEM 131, which is a three credit course, the responses to (1) and (2) are appropriately adjusted.)

Question 22 on p. 374 asks the student if:
The Study Guides should be: (1) more specific (2) about the same or (3) more general.

The last item displayed in these bar graphs is No. 22 which asks what percent of a student's time is spent studying from 'old' exams. The height of each category of percent-time-spent is the percentage of students responding in that category. The quartiles of the distribution are indicated by triangles along the bottom edge of each graph. Bars are of alternating design merely to facilitate interpretation.

F.3. Item means and correlations with course grade

Selected items of interest from the precourse and postcourse attitude surveys and the course evaluation form are listed in Table F.3. Each entry includes the item mean on a scale of 1.0 to 5.0, the standard deviation of the responses, the correlation between the student's response and his grade, and the mean grade for all students who

responded to the item. The grade means are different for different items because the same students did not respond to all the same items; incomplete survey return leads to a difference between the grade mean for an item and the grade mean for the total enrollment in the course.

The first fourteen variables are from the precourse attitude survey and are labeled '14-PRE' through '38' with these labels corresponding to the numbers of the survey questions on the sample form shown in Figure 4.6, p. 218. The second fourteen variables are from the postcourse survey and are labeled '14-POST' through '38' to correspond to the question numbers in Figure 4.7 on p. 219. The last eight variables are questions from the course evaluation form and are labeled '1' through '14' in agreement with Figure 4.4 on p. 214.

These data are the output of computer program Minerva. The means from these tables are slightly different from some means quoted in Chapter IV or given in Table F.1 because the numbers of students on which these means are based are different. In order to calculate correlations with course grade, only those surveys which included a student number could be used, and several percent of each term's surveys are returned without student number identification. These 'anonomous' responses are counted in the overall means and response frequency results but not in Tables F.3 or F.4.

F.4. Factor means and correlations with course grade

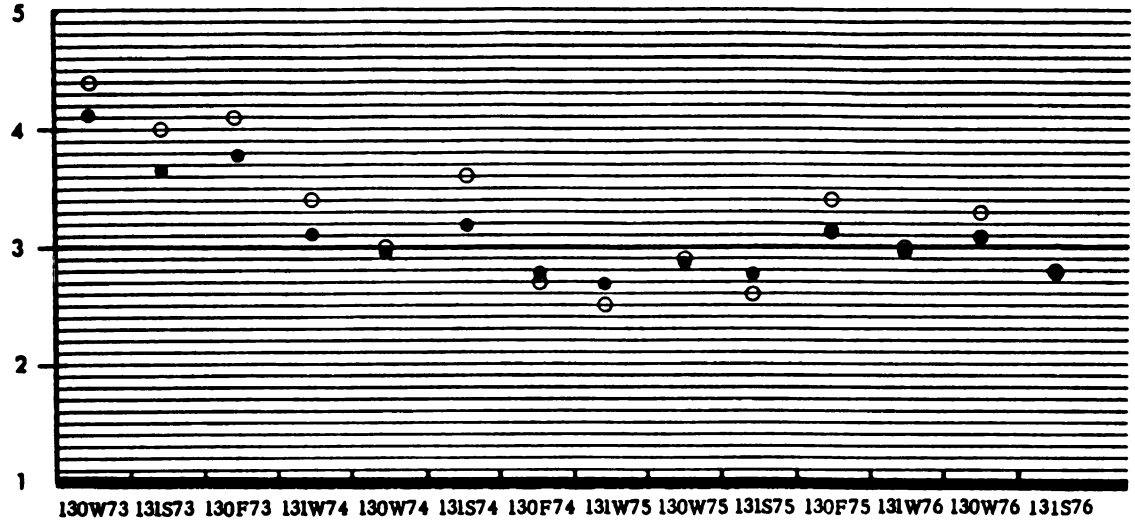
The survey questions from Table F.3 are grouped into factors and the factor mean and factor correlation with course grade calculated for this table. The survey items which make up these twelve factors are listed here, taken from Table 4.2, p. 232.

Factor 1	Nos. 14, 19, 21 - precourse attitude survey
Factor 2	Nos. 22, 23, 30 - precourse attitude survey
Factor 3	Nos. 31, 35, 36, 40 - precourse attitude survey
Factor 4	Nos. 32, 34, 37, 39 - precourse attitude survey
Factor 5	Nos. 33, 38 - precourse attitude survey
Factor 6	Nos. 14, 19, 21 - postcourse attitude survey
Factor 7	Nos. 22, 23, 30 - postcourse attitude survey
Factor 8	Nos. 31, 35, 36, 40 - postcourse attitude survey
Factor 9	Nos. 32, 34, 37, 39 - postcourse attitude survey
Factor 10	Nos. 33, 38 - postcourse attitude survey
Factor 11	Nos. 1, 2, 3, 10, 14 - course evaluation form
Factor 12	Nos. 4, 5, 6 - course evaluation form

The means listed in this table may differ from means based on all survey responses because any surveys submitted without student number identification could not be used to calculate a correlation with course grade and were not included in the calculation of the mean.

Table F.1 Fourteen graphs of average response to Items 1 through 14 from the course evaluation form indicating mean [●] and interpolated median [○] responses for each of fourteen terms with major enrollments

1. How much did you like this method of teaching a course compared to the traditional lecture/recitation method?



2. How much did you learn under this method compared to the amount you might have learned under the traditional method?

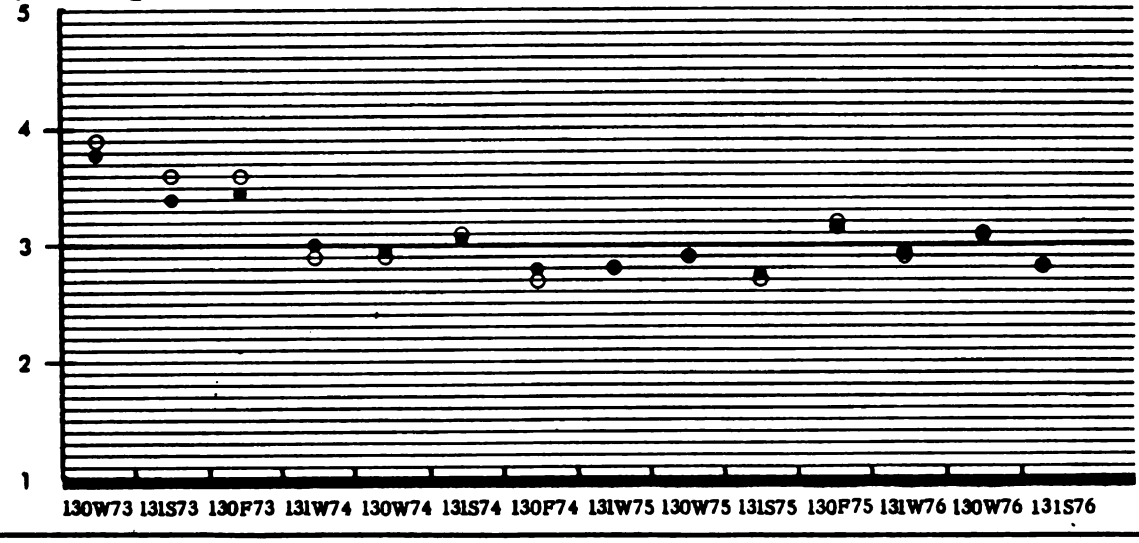
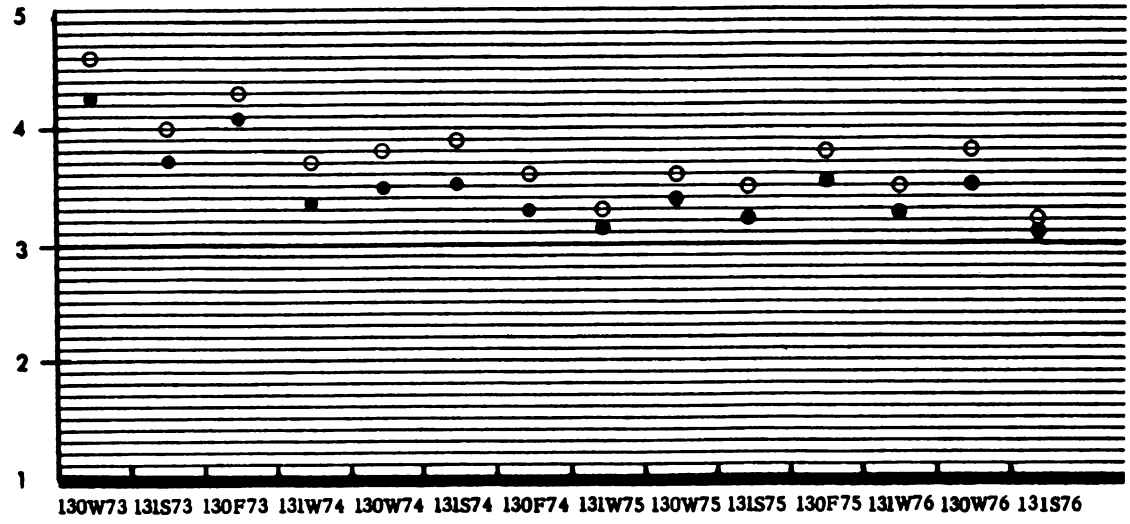
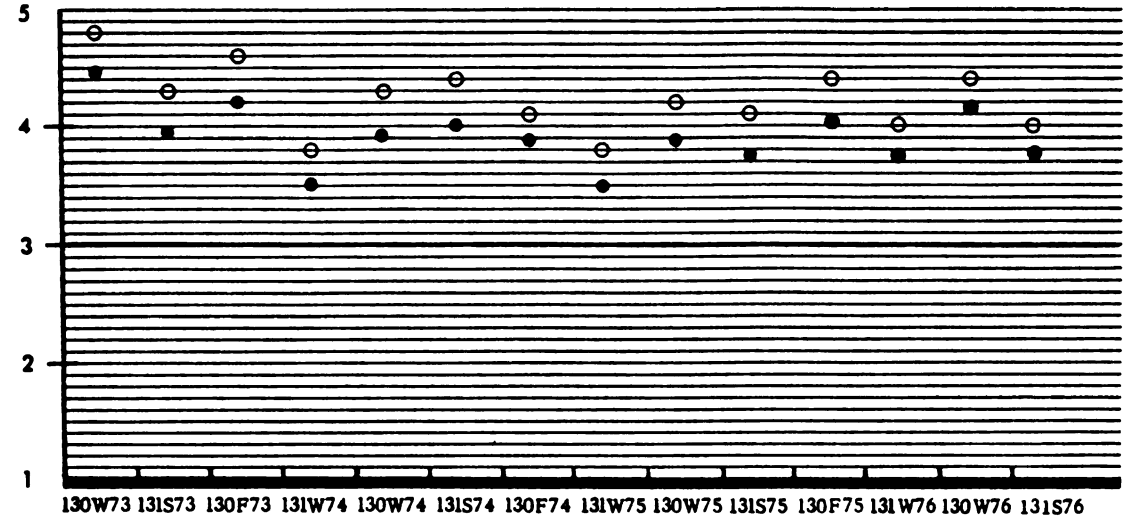


Table F.1 (cont'd)

3. I liked the self pacing.



4. I liked the exam procedure.



5. I liked how the final exam counted toward my final grade.

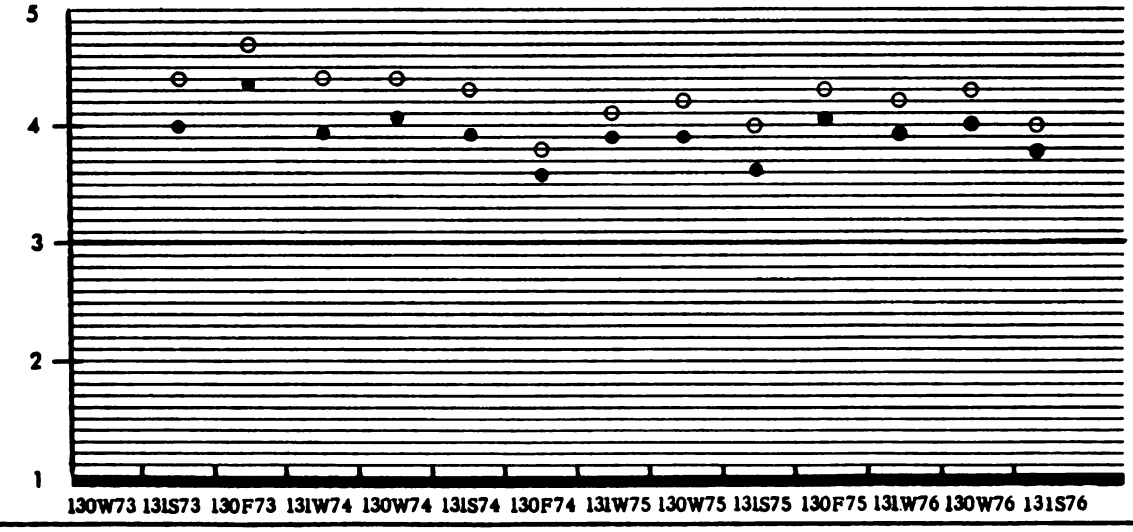
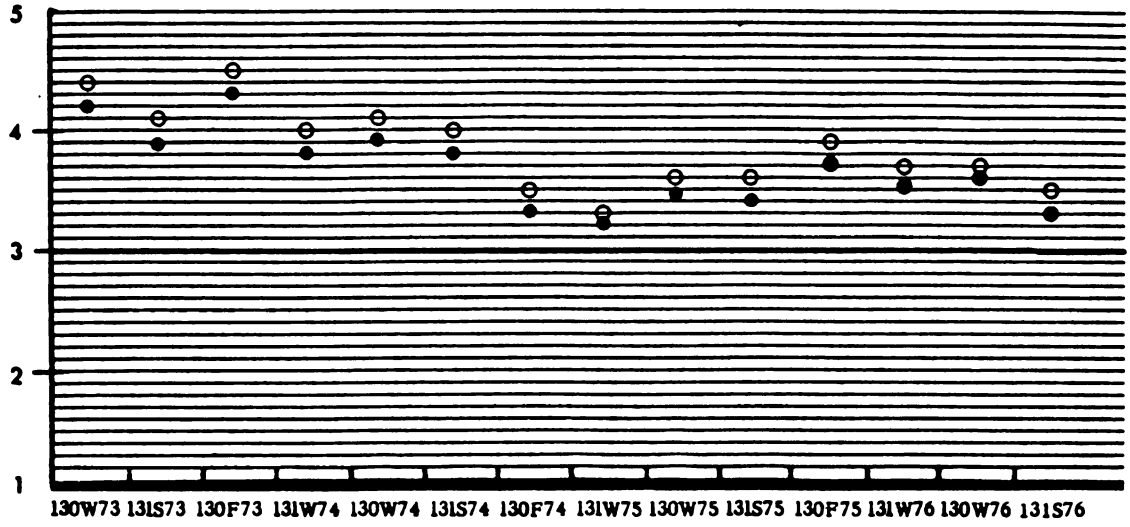
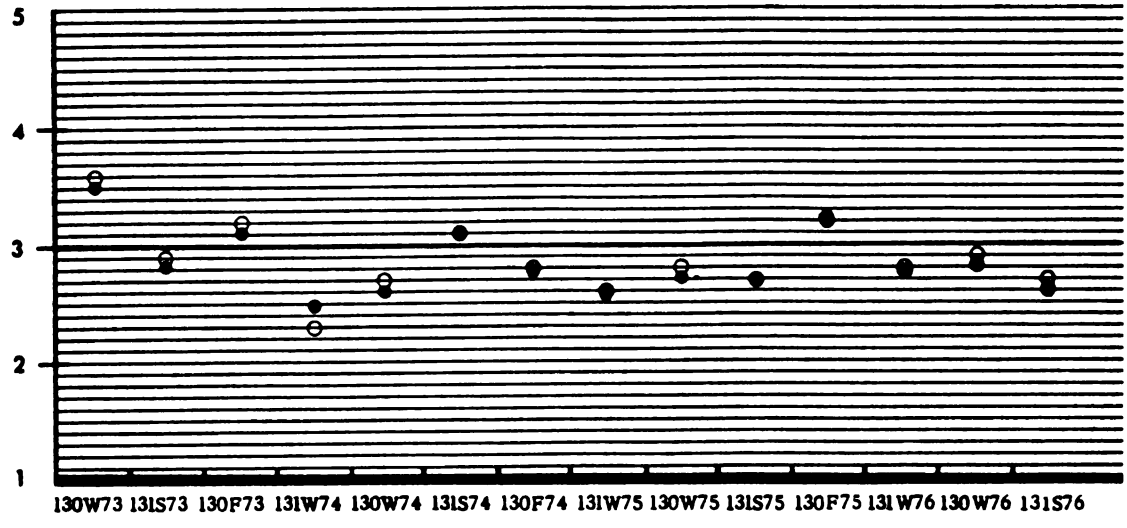


Table F.1 (cont'd)

6. I liked the grading scale.



7. I liked the CEM Room.



8. I liked the textbook.

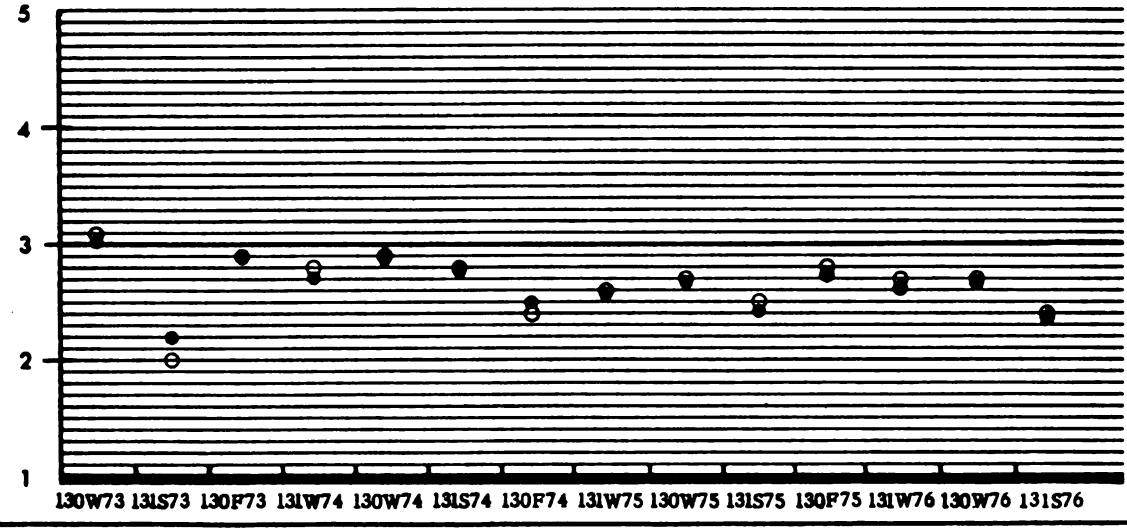
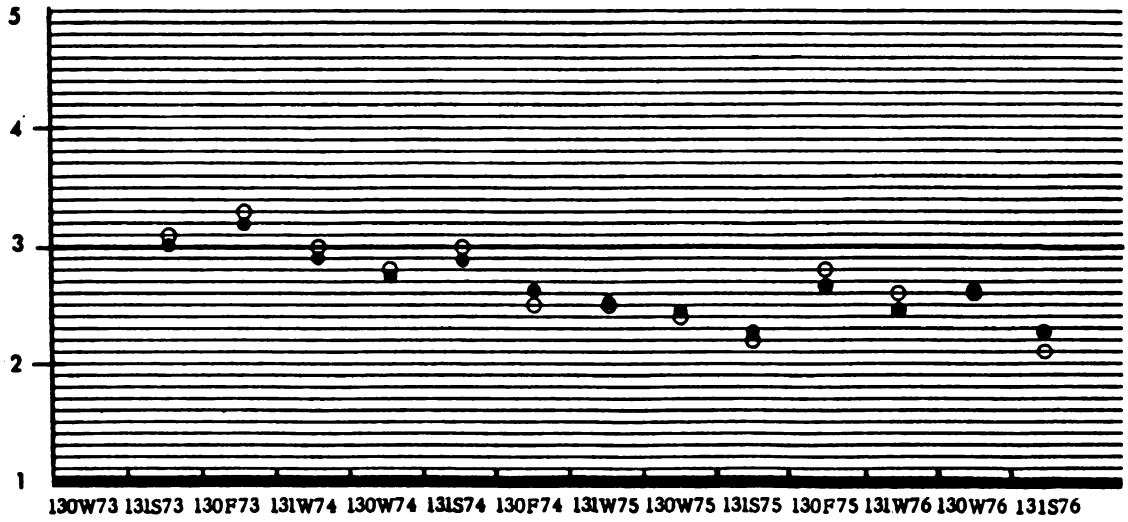
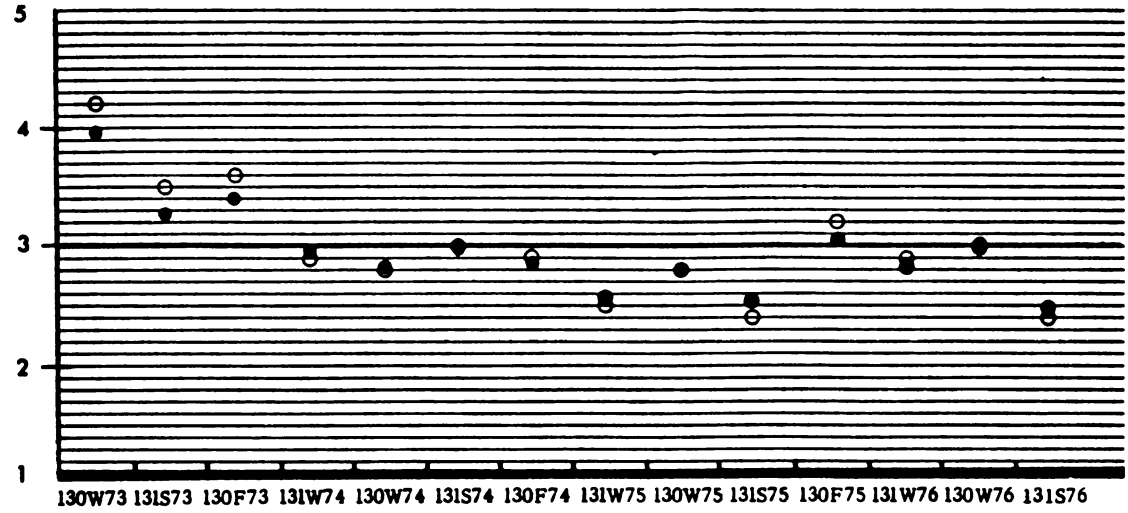


Table F.1 (cont'd)

9. I liked the Tapes.



10. The audio tapes are better than lectures would have been.



11. After I did what the study guide said, I felt prepared to take an exam.

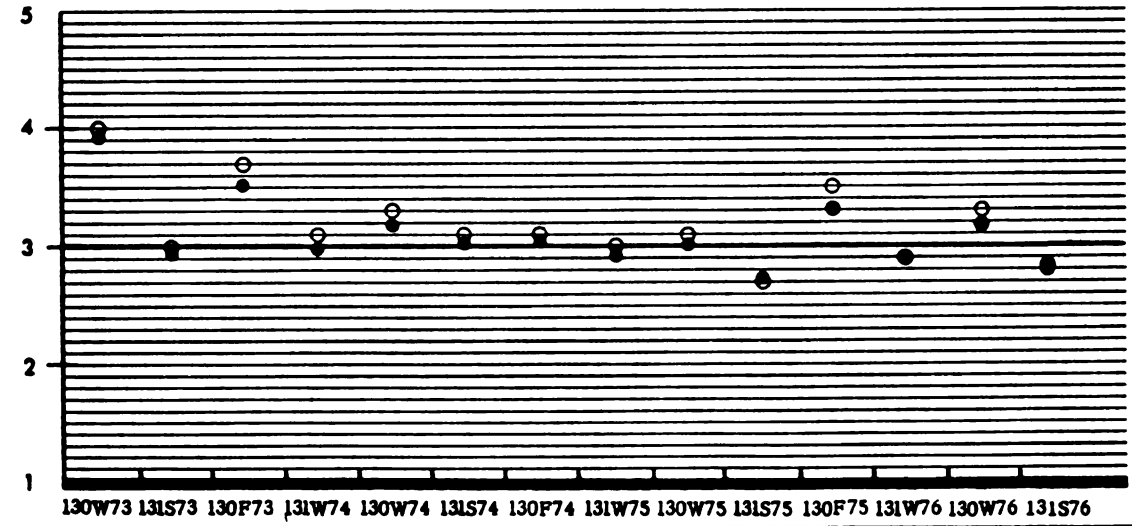
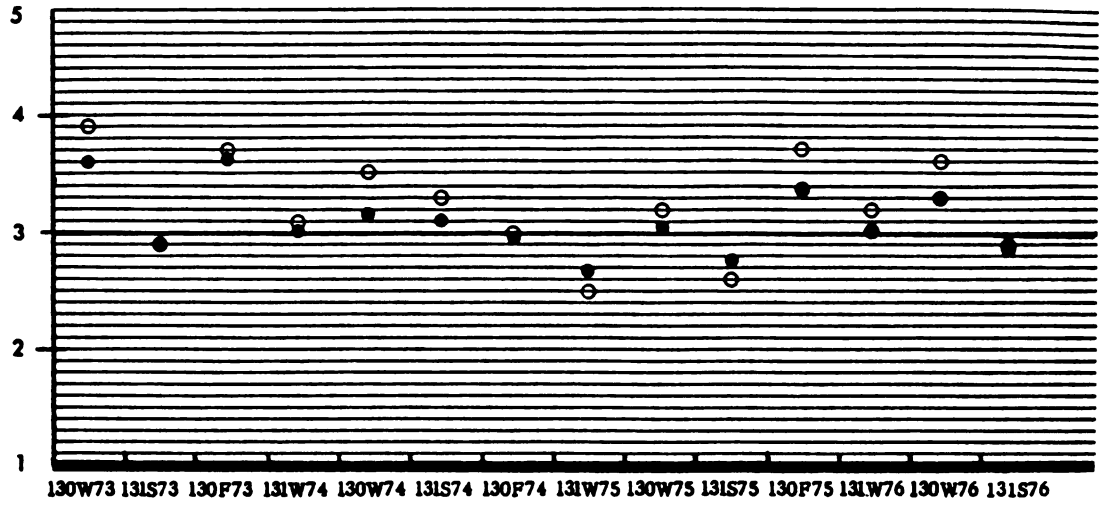
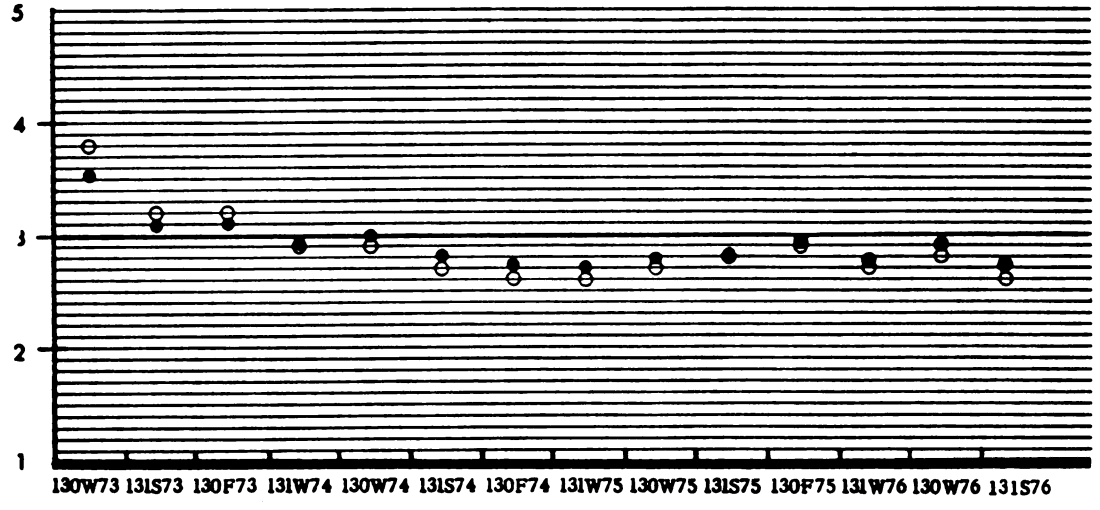


Table F.1 (cont'd)

12. The grade I get reflects the amount of work I put into the course.



13. I did not feel nervous or anxious about taking an exam.



14. If I were given the choice, I would choose this method for a course rather than the lecture/recitation-three-exams-and-a-final method.

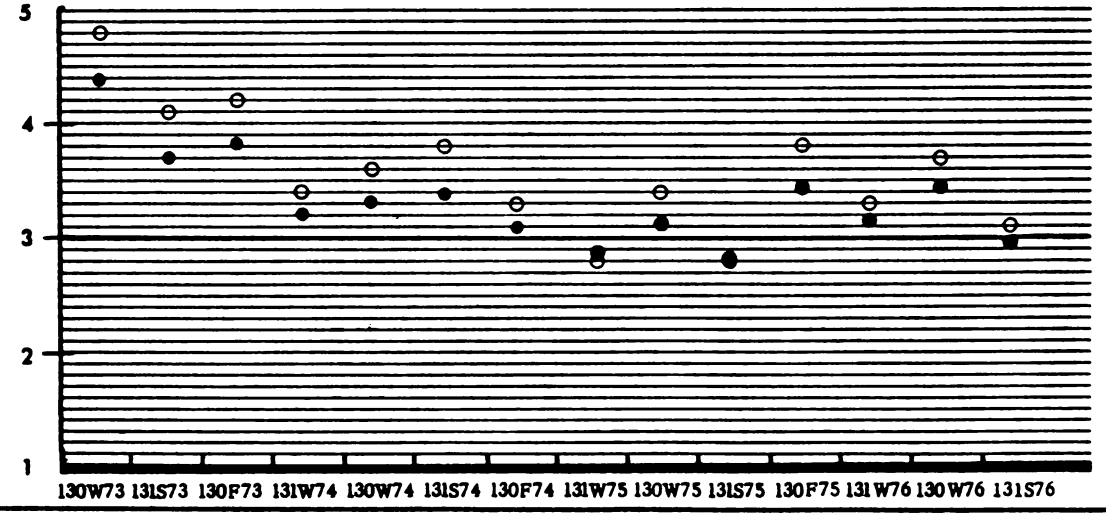
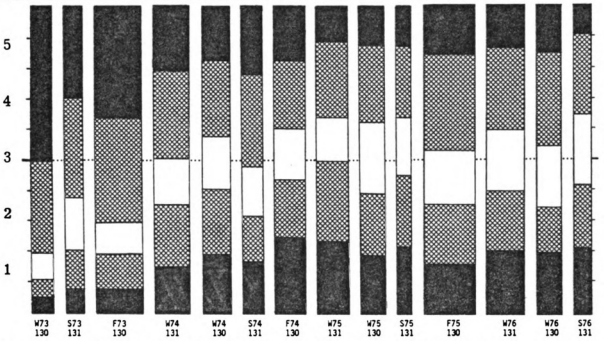
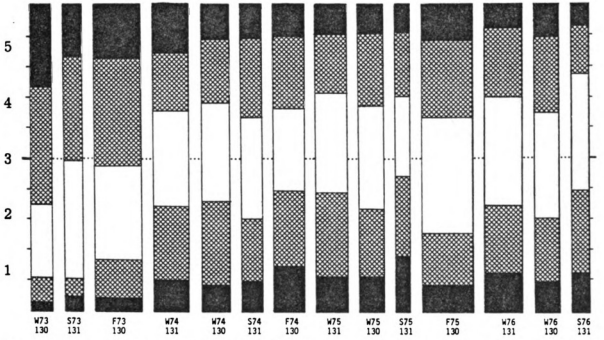


Table F.2 Twenty-four bar graphs of percent response in each response category for Items 1 through 14 and 18 through 22 for terms with major enrollments

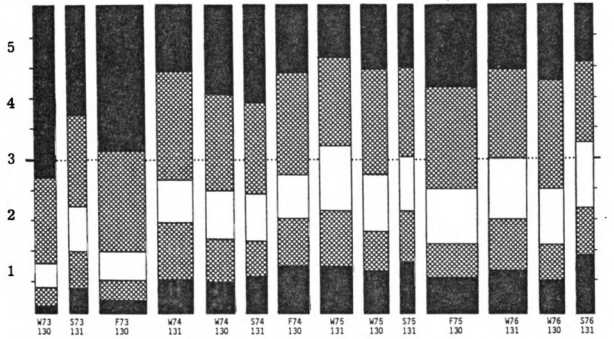


1. How much did you like this method of teaching a course compared to the traditional lecture/recitation method?

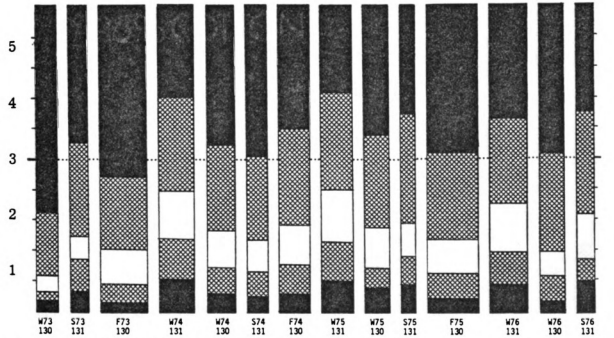


2. How much did you learn under this method compared to the amount you might have learned under the traditional method?

Table F.2 (cont'd)

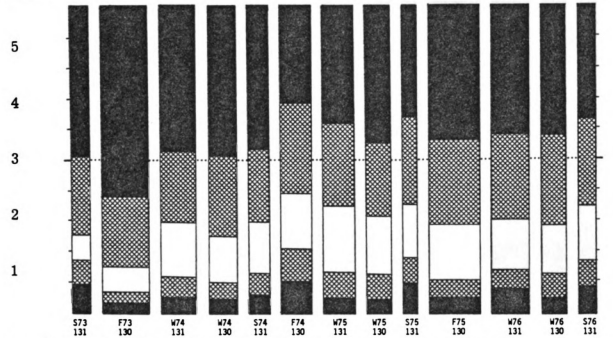


3. I liked the self pacing.

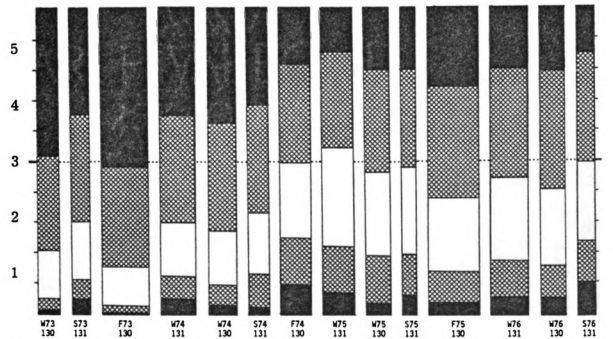


4. I liked the exam procedure

Table F.2 (cont'd)

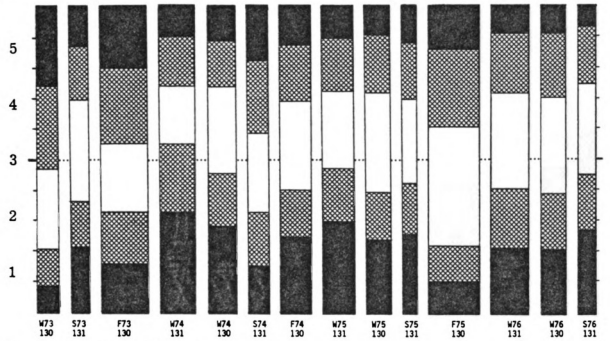


5. I liked how the final exam counted toward my final grade.

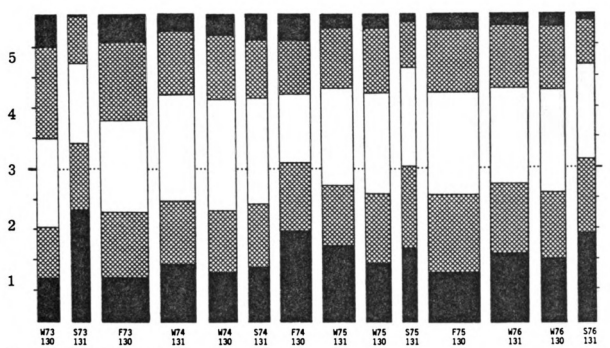


6. I liked the grading scale.

Table F.2 (cont'd)

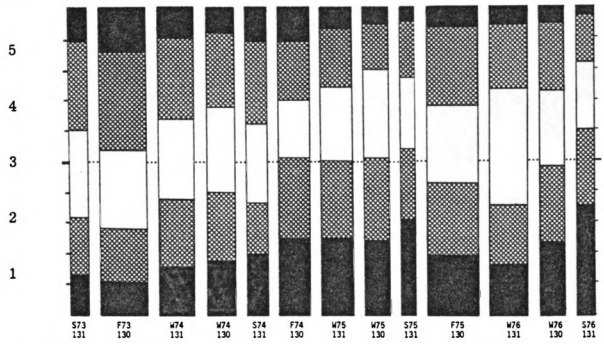


7. I liked the CEM Room.

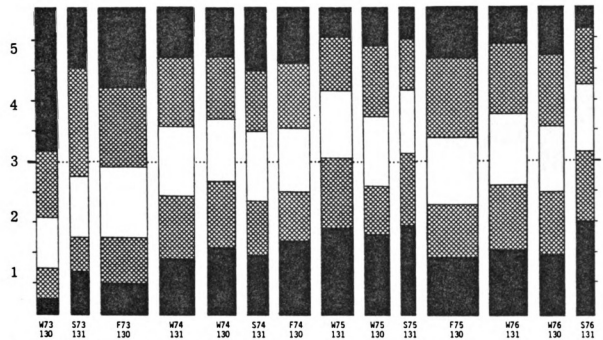


8. I liked the textbook.

Table F.2 (cont'd)

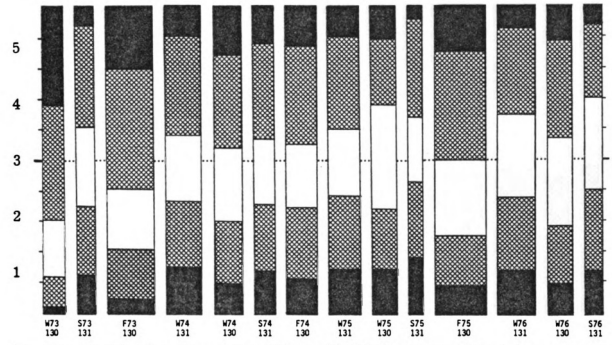


9. I liked the Tapes.

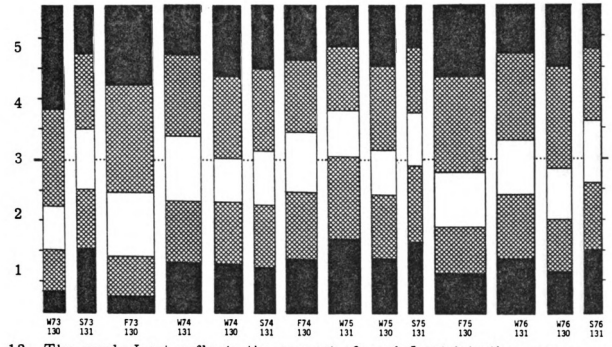


10. The audio tapes are better than lectures would have been.

Table F.2 (cont'd)

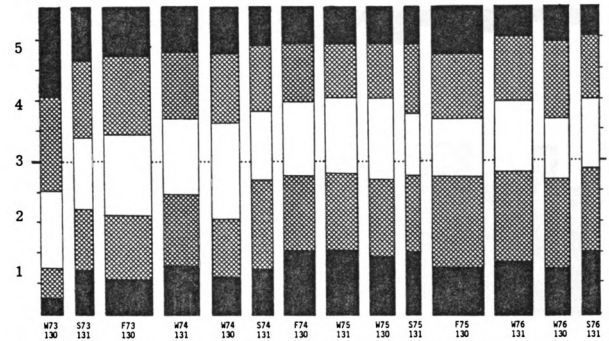


11. After I did what the study guide said, I felt prepared to take an exam.

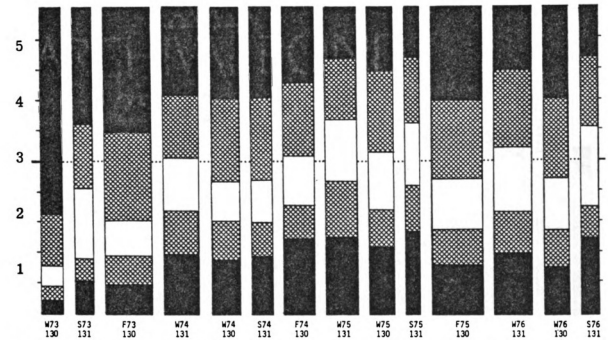


12. The grade I get reflects the amount of work I put into the course.

Table F.2 (cont'd)

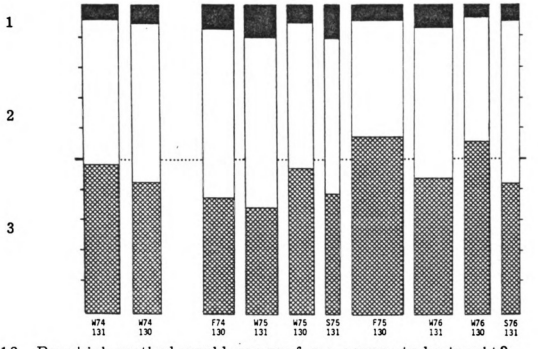


13. I did not feel nervous or anxious about taking an exam.

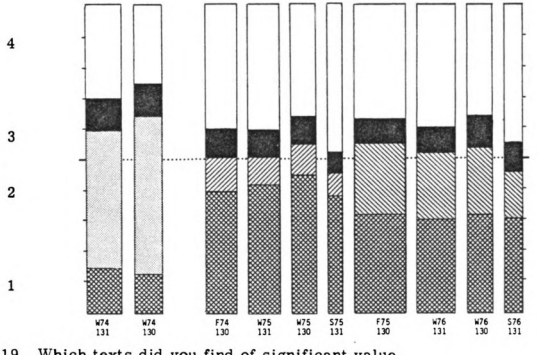


14. If I were given the choice, I would choose this method for a course rather than the lecture/recitation-three-exams-and-a-final method.

Table F.2 (cont'd)

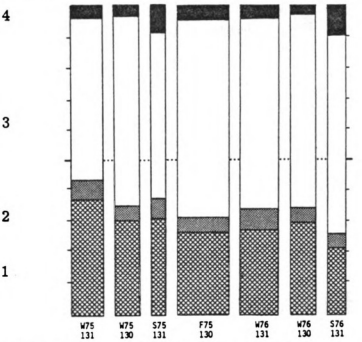


18. By which method would you prefer a course to be taught?

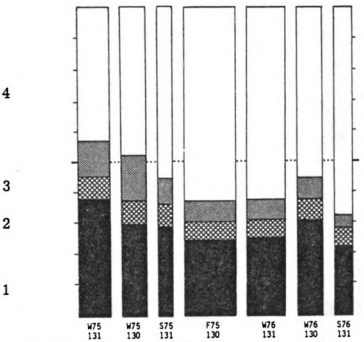


19. Which texts did you find of significant value.

Table F.2 (cont'd)

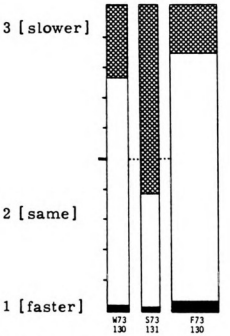


20. How did you most often listen to audio tapes?

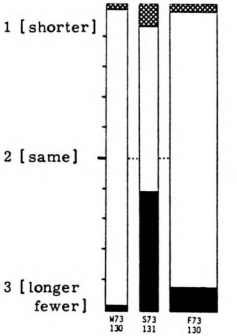


21. How many audio tapes did you duplicate?

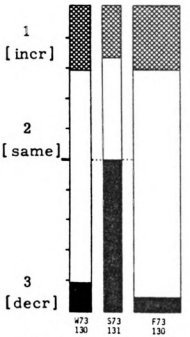
Table F.2 (cont'd)



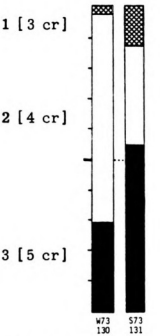
18. The pacing in the course should be...



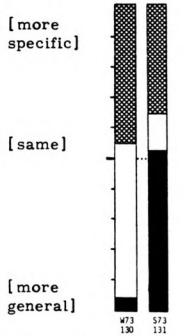
19. The exams should be...



20. The number of Study Questions should be...



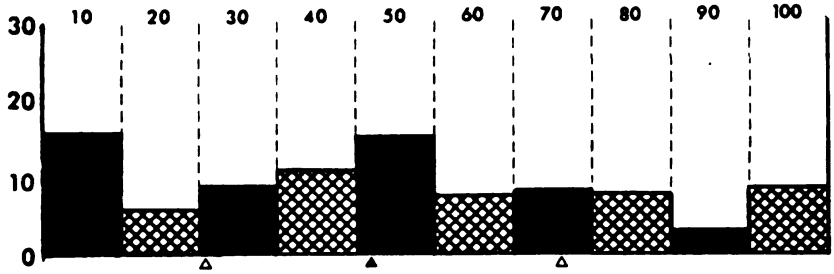
21. This course should be...



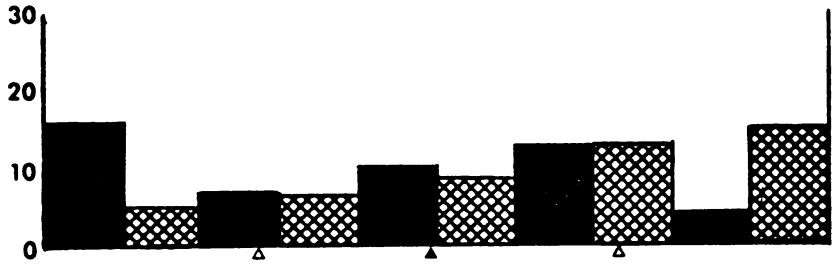
22. The Study Guides should be...

Table F.2 (cont'd)

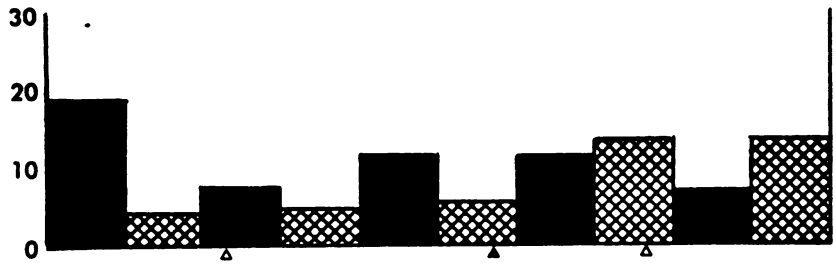
CEM131S74



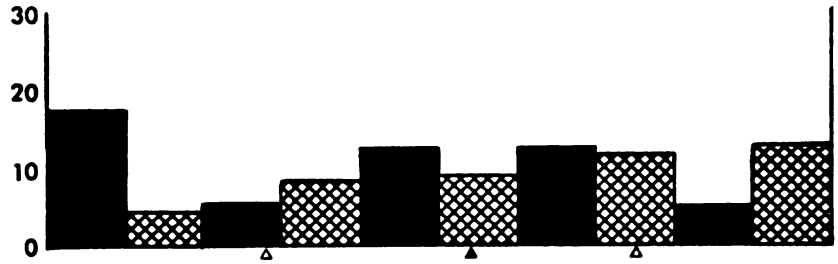
CEM130F74



CEM131W75



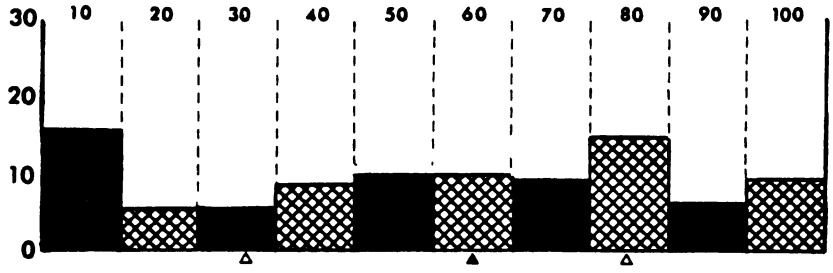
CEM130W75



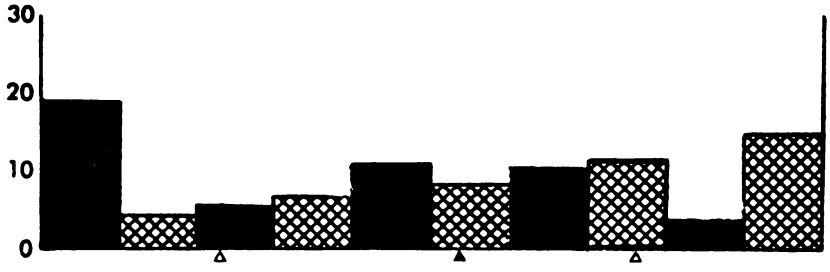
22. What percentage of your effort did you expend studying from "old" exams?

Table F.2 (cont'd)

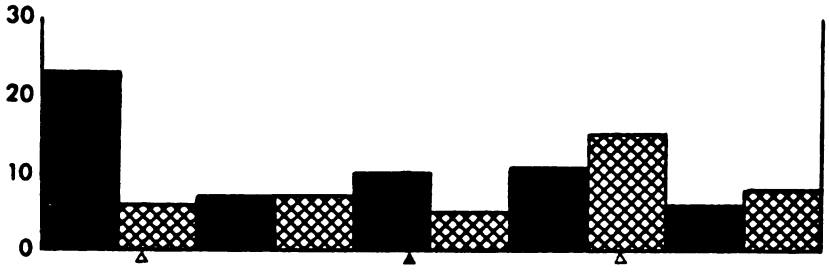
CEM131S75



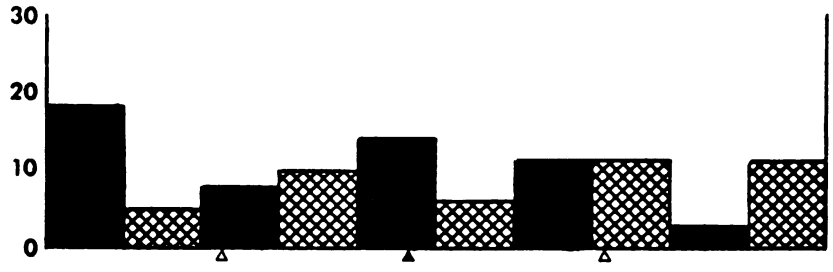
CEM130F75



CEM131W76



CEM130W76



CEM131S76

Table F.3.1 Item means and correlations with course grade

WINTER 1973 CHEM 130 N = 556

VARIABLE	NUMBER OF RECORDS	MEAN	STANDARD DEVIATION	CORRELATION WITH GRADE	GRADE MEAN
14-PRE	466.	2.8627	.8431	-.1170	2.7672
19	468.	2.5855	1.2639	.0151	2.7671
21	471.	1.9172	.7924	-.0922	2.7675
22	468.	3.8632	.8858	.0317	2.7692
23	469.	3.5693	1.0878	-.0081	2.7655
30	468.	3.5000	.9860	.0284	2.7682
31,35	459.	4.2580	.7712	-.0263	2.7767
32,34	461.	3.0700	.6629	.0725	2.7755
33	462.	3.3874	.8993	.0228	2.7662
36,40	447.	4.3378	.7026	-.0171	2.7707
37,39	448.	3.0242	.6234	.0725	2.7656
38	449.	3.1938	.9006	-.0246	2.7673
14-POST	482.	2.6203	.9764	-.3162	2.8278
19	474.	2.3966	1.2025	.0992	2.8259
21	477.	1.9644	.8497	-.2859	2.8270
22	472.	3.4873	1.0576	.1697	2.8231
23	473.	2.9852	1.1457	.1528	2.8256
30	457.	3.0088	.9989	.1802	2.8359
31,35	408.	4.1120	.8622	.1238	2.8419
32,34	428.	2.8184	.7431	-.0099	2.8318
33	455.	3.5736	.8985	.2568	2.8231
36,40	344.	4.2924	.7587	.1384	2.8648
37,39	365.	2.8521	.6336	-.1446	2.8493
38	406.	3.3202	.8743	.0365	2.7968
1	442.	4.1403	1.1148	.3482	2.8416
2	435.	3.7908	1.0171	.3058	2.8425
3	457.	4.2451	1.0361	.3316	2.8359
4	456.	4.4583	.9724	.1919	2.8355
5	452.	2.9845	1.2000	.0775	2.8341
6	455.	4.2066	.9406	.2090	2.8352
10	458.	2.0961	1.2571	-.1522	2.8373
14	458.	1.6878	1.2134	-.3122	2.8373

Table F.3.2 Item means and correlations with course grade

SPRING 1973 CHEM 131 N = 498

VARIABLE	NUMBER OF RECORDS	MEAN	STANDARD DEVIATION	CORRELATION WITH GRADE	GRADE MEAN
14-PRE	0.	.5000	.5000	.0000	.5000
19	0.	.5000	.5000	.0000	.5000
21	0.	.5000	.5000	.0000	.5000
22	0.	.5000	.5000	.0000	.5000
23	0.	.5000	.5000	.0000	.5000
30	0.	.5000	.5000	.0000	.5000
31,35	0.	.5000	.5000	.0000	.5000
32,34	0.	.5000	.5000	.0000	.5000
33	0.	.5000	.5000	.0000	.5000
36,40	0.	.5000	.5000	.0000	.5000
37,39	0.	.5000	.5000	.0000	.5000
38	0.	.5000	.5000	.0000	.5000
14-POST	386.	3.0026	1.0369	-.3289	2.4767
19	385.	2.0623	1.1447	.1618	2.4740
21	385.	2.6286	1.1257	-.3171	2.4753
22	386.	3.0440	1.1177	.2225	2.4767
23	385.	2.5714	1.0596	.2070	2.4766
30	380.	2.6763	.9857	.2946	2.4645
31,35	372.	4.2519	.8154	.1000	2.4664
32,34	375.	2.9230	.7412	-.0913	2.4587
33	376.	3.3457	1.0276	.2444	2.4614
36,40	350.	4.3959	.6428	.1035	2.4729
37,39	355.	2.8849	.6563	-.1173	2.4592
38	361.	3.1496	.8868	.0983	2.4668
1	408.	3.6691	1.2856	.3377	2.4865
2	403.	3.4020	1.1215	.4052	2.4814
3	422.	3.7417	1.2813	.3832	2.4716
4	423.	3.9811	1.2164	.2135	2.4752
5	424.	3.9953	1.2922	.0990	2.4741
6	421.	3.9026	1.1060	.1476	2.4822
10	426.	2.7347	1.3523	-.1204	2.4754
14	426.	2.2746	1.4199	-.2705	2.4754

Table F.3.3 Item means and correlations with course grade

FALL 1973 CHEM 130 N = 1221

VARIABLE	NUMBER OF RECORDS	MEAN	STANDARD DEVIATION	CORRELATION WITH GRADE	GRADE MEAN
14-PRE	1190.	2.7509	.8923	-.0947	3.1200
19	1193.	2.9646	1.2688	-.0378	3.1192
21	1185.	1.9629	.7760	-.0197	3.1172
22	1185.	3.9357	.9420	.0524	3.1167
23	1196.	3.5748	1.0472	-.0212	3.1177
30	1184.	3.5063	.9469	.0349	3.1150
31,35	1195.	4.3670	.7110	.0210	3.1146
32,34	1194.	3.1118	.6505	.0725	3.1124
33	1180.	3.4255	.9249	.0190	3.1159
36,40	1164.	4.4092	.6768	-.0003	3.1151
37,39	1167.	3.0908	.5916	.0070	3.1167
38	1175.	3.2214	.8836	-.0451	3.1191
14-POST	994.	2.5674	.9345	-.2605	3.1766
19	988.	2.5860	1.2147	.0377	3.1781
21	994.	2.1177	.9089	-.2279	3.1766
22	992.	3.5474	1.0672	.2367	3.1749
23	992.	2.9244	1.0838	.1837	3.1769
30	992.	3.0887	.9663	.2152	3.1759
31,35	974.	4.2297	.8195	.0199	3.1735
32,34	985.	2.9991	.7136	.0226	3.1766
33	987.	3.4671	.9788	.2512	3.1758
36,40	920.	4.3571	.6954	.0938	3.1777
37,39	943.	2.9555	.6216	-.0685	3.1702
38	943.	3.2206	.8763	.0106	3.1702
1	0.	.5000	.5000	.0000	.5000
2	0.	.5000	.5000	.0000	.5000
3	0.	.5000	.5000	.0000	.5000
4	0.	.5000	.5000	.0000	.5000
5	0.	.5000	.5000	.0000	.5000
6	0.	.5000	.5000	.0000	.5000
10	0.	.5000	.5000	.0000	.5000
14	0.	.5000	.5000	.0000	.5000

Table F.3.4 Item means and correlations with course grade

WINTER 1974 CHEM 131 N = 1110

VARIABLE	NUMBER OF RECORDS	MEAN	STANDARD DEVIATION	CORRELATION WITH GRADE	GRADE MEAN
14-PRE	0.	.5000	.5000	.0000	.5000
19	0.	.5000	.5000	.0000	.5000
21	0.	.5000	.5000	.0000	.5000
22	0.	.5000	.5000	.0000	.5000
23	0.	.5000	.5000	.0000	.5000
30	0.	.5000	.5000	.0000	.5000
31,35	0.	.5000	.5000	.0000	.5000
32,34	0.	.5000	.5000	.0000	.5000
33	0.	.5000	.5000	.0000	.5000
36,40	0.	.5000	.5000	.0000	.5000
37,39	0.	.5000	.5000	.0000	.5000
38	0.	.5000	.5000	.0000	.5000
14-POST	475.	2.9011	1.0202	-.3335	2.5789
19	467.	2.1542	1.1549	.1350	2.5664
21	472.	2.7246	1.1182	-.3105	2.5710
22	470.	3.0702	1.0706	.2653	2.5745
23	470.	2.5809	1.1113	.2658	2.5660
30	463.	2.6998	1.0217	.2680	2.5734
31,35	403.	4.2740	.8470	-.0341	2.5881
32,34	418.	3.1135	.7562	-.0125	2.5706
33	451.	3.1996	1.0466	.3408	2.5732
36,40	310.	4.3272	.7918	-.0005	2.6145
37,39	371.	3.0716	.6680	-.0887	2.5741
38	396.	3.1263	.9397	.0729	2.5985
1	713.	3.1697	1.3732	.3336	2.6241
2	714.	2.9916	1.1905	.3624	2.6225
3	715.	3.3748	1.3071	.2807	2.6182
4	725.	3.5683	1.3328	.1659	2.6276
5	724.	4.0110	1.2382	.1401	2.6257
6	725.	3.8938	1.1159	.1426	2.6290
10	730.	3.0411	1.3215	-.2152	2.6295
14	730.	2.7945	1.5248	-.2810	2.6295

Table F.3.5 Item means and correlations with course grade

WINTER 1974 CHEM 130 N = 780

VARIABLE	NUMBER OF RECORDS	MEAN	STANDARD DEVIATION	CORRELATION WITH GRADE	GRADE MEAN
14-PRE	106.	3.2075	1.0067	-.0386	2.8160
19	66.	2.4242	1.3603	-.0769	2.7348
21	108.	2.2130	.9235	-.0763	2.8009
22	103.	3.8835	1.0173	-.0627	2.7864
23	92.	3.3370	1.0962	-.0827	2.7826
30	114.	3.3509	.9732	-.0504	2.8158
31,35	83.	4.4148	.6456	.0896	2.7530
32,34	99.	3.0433	.6502	.1371	2.8131
33	106.	3.3585	.8816	.1176	2.7972
36,40	91.	4.5228	.5699	.2489	2.7582
37,39	104.	3.0549	.6113	-.0247	2.7788
38	109.	3.4128	.7688	.1990	2.8119
14-POST	0.	.5000	.5000	.0000	.5000
19	0.	.5000	.5000	.0000	.5000
21	0.	.5000	.5000	.0000	.5000
22	0.	.5000	.5000	.0000	.5000
23	0.	.5000	.5000	.0000	.5000
30	0.	.5000	.5000	.0000	.5000
31,35	0.	.5000	.5000	.0000	.5000
32,34	0.	.5000	.5000	.0000	.5000
33	0.	.5000	.5000	.0000	.5000
36=40	0.	.5000	.5000	.0000	.5000
37,39	0.	.5000	.5000	.0000	.5000
38	0.	.5000	.5000	.0000	.5000
1	560.	3.0000	1.3823	.2534	2.8161
2	544.	2.9614	1.1633	.2638	2.8125
3	571.	3.5306	1.3129	.2304	2.8152
4	573.	3.9913	1.1970	.0516	2.8141
5	570.	4.0702	1.1571	.1902	2.8184
6	573.	3.9494	1.0532	.0860	2.8150
10	576.	3.2031	1.3777	-.1531	2.8134
14	576.	2.7743	1.5990	-.1979	2.8134

Table F.3.6 Item means and correlations with course grade

SPRING 1974 CHEM 131 N = 660

VARIABLE	NUMBER OF RECORDS	MEAN	STANDARD DEVIATION	CORRELATION WITH GRADE	GRADE MEAN
14-PRE	0.	.5000	.5000	.0000	.5000
19	0.	.5000	.5000	.0000	.5000
21	0.	.5000	.5000	.0000	.5000
22	0.	.5000	.5000	.0000	.5000
23	0.	.5000	.5000	.0000	.5000
30	0.	.5000	.5000	.0000	.5000
31,35	0.	.5000	.5000	.0000	.5000
32,34	0.	.5000	.5000	.0000	.5000
33	0.	.5000	.5000	.0000	.5000
36,40	0.	.5000	.5000	.0000	.5000
37,39	0.	.5000	.5000	.0000	.5000
38	0.	.5000	.5000	.0000	.5000
14-POST	441.	3.0499	1.0173	-.3837	2.5200
19	398.	2.0402	1.0743	.1101	2.5239
21	398.	2.6508	1.1369	-.3797	2.5239
22	398.	3.0352	1.0745	.3238	2.5239
23	398.	2.5327	1.0879	.2726	2.5239
30	395.	2.6000	.9096	.3325	2.5215
31,35	392.	4.1385	.8414	-.0112	2.5255
32,34	390.	2.9678	.7435	-.0210	2.5231
33	393.	3.3384	1.0982	.3333	2.5267
36,40	372.	4.2919	.7792	.0480	2.5202
37,39	372.	2.9662	.6338	-.0426	2.5148
38	375.	3.1920	.8977	.1317	2.5333
1	433.	3.2032	1.4322	.3385	2.5185
2	428.	3.0444	1.1623	.3659	2.5199
3	443.	3.5621	1.3762	.3188	2.5293
4	442.	4.0407	1.1760	.1128	2.5339
5	441.	3.9637	1.2376	.2028	2.5317
6	443.	3.8668	1.0678	.0978	2.5293
10	443.	3.0045	1.3884	-.1677	2.5293
14	443.	2.6524	1.5058	-.2933	2.5293

Table F.3.7 Item means and correlations with course grade

FALL 1974 CHEM 130 N = 1242

VARIABLE	NUMBER OF RECORDS	MEAN	STANDARD DEVIATION	CORRELATION WITH GRADE	GRADE MEAN
14-PRE	1088.	2.7325	.8855	-.0836	2.6324
19	1129.	2.7883	1.2579	.0970	2.6262
21	1121.	2.0241	.8094	-.0456	2.6231
22	1121.	4.0134	.9468	.0586	2.6285
23	1125.	3.5422	1.0796	-.0437	2.6289
30	1119.	3.6077	.9700	.0239	2.6206
31,35	1121.	4.4179	.6993	-.0057	2.6204
32,34	1128.	3.2143	.6553	.0929	2.6215
33	1128.	3.3963	.9368	.0440	2.6210
36,40	1104.	4.4798	.6047	.0045	2.6232
37,39	1111.	3.1458	.5767	-.0127	2.6224
38	1111.	3.1665	.8971	-.0432	2.6211
14-POST	379.	2.8865	1.0678	-.2833	2.5844
19	368.	2.3940	1.1886	.1067	2.5802
21	371.	2.6550	1.1561	-.3629	2.5782
22	372.	3.2043	1.0980	.3183	2.5793
23	368.	2.6060	1.2112	.2343	2.5747
30	366.	2.7432	1.0426	.2813	2.5820
31,35	356.	4.2665	.8493	-.0682	2.5758
32,34	357.	3.1673	.7714	-.0183	2.5770
33	362.	3.2210	1.1325	.3446	2.5760
36,40	347.	4.3380	.7453	.0240	2.5807
37,39	349.	3.1465	.6166	-.1085	2.5788
38	349.	3.0774	.9408	.1090	2.5788
1	556.	2.7896	1.4081	.2996	2.5944
2	550.	2.7909	1.1623	.2986	2.6018
3	582.	3.3299	1.3576	.2930	2.6040
4	582.	3.9502	1.1924	.1209	2.6057
5	583.	3.5901	1.3051	.1055	2.6055
6	581.	3.3460	1.1977	.0077	2.6041
10	584.	3.1387	1.4165	-.1982	2.6019
14	584.	2.9264	1.5271	-.2633	2.6019

Table F.3.8 Item means and correlations with course grade

WINTER 1975 CHEM 131 N = 1112

VARIABLE	NUMBER OF RECORDS	MEAN	STANDARD DEVIATION	CORRELATION WITH GRADE	GRADE MEAN
14-PRE	0.	.5000	.5000	.0000	.5000
19	0.	.5000	.5000	.0000	.5000
21	0.	.5000	.5000	.0000	.5000
22	0.	.5000	.5000	.0000	.5000
23	0.	.5000	.5000	.0000	.5000
30	0.	.5000	.5000	.0000	.5000
31,35	0.	.5000	.5000	.0000	.5000
32,34	0.	.5000	.5000	.0000	.5000
33	0.	.5000	.5000	.0000	.5000
36,40	0.	.5000	.5000	.0000	.5000
37,39	0.	.5000	.5000	.0000	.5000
38	0.	.5000	.5000	.0000	.5000
14-POST	437.	2.9794	1.0511	-.2723	2.5584
19	430.	2.1163	1.1596	.1196	2.5581
21	432.	2.7315	1.1394	-.2611	2.5579
22	433.	3.0739	1.0612	.2449	2.5543
23	433.	2.5381	1.1184	.1836	2.5543
30	425.	2.6612	.9908	.2625	2.5506
31,35	417.	4.2326	.7984	-.0439	2.5588
32,34	417.	3.1816	.7319	-.0432	2.5468
33	420.	3.1667	1.0672	.2897	2.5524
36,40	406.	4.3540	.7363	.0858	2.5591
37,39	405.	3.1108	.6311	-.0774	2.5580
38	406.	3.0764	.9784	.0634	2.5591
1	527.	2.7438	1.3172	.3536	2.5626
2	525.	2.8267	1.0684	.3844	2.5657
3	550.	3.1680	1.3093	.3326	2.5682
4	553.	3.5371	1.3336	.0874	2.5660
5	549.	3.8652	1.1889	.1760	2.5701
6	549.	3.2732	1.1170	.0517	2.5592
10	562.	3.0712	1.3819	-.0720	2.5703
14	562.	3.2085	1.4909	-.0731	2.5703

Table F.3.9 Item means and correlations with course grade

WINTER 1975 CHEM 130 N = 634

VARIABLE	NUMBER OF RECORDS	MEAN	STANDARD DEVIATION	CORRELATION WITH GRADE	GRADE MEAN
14-PRE	449.	3.0156	.8610	-.0631	2.5624
19	450.	2.4178	1.1846	.0388	2.5644
21	452.	2.3650	.9533	-.0207	2.5597
22	448.	3.6629	1.0178	.0220	2.5625
23	446.	3.2713	1.1107	-.0225	2.5527
30	447.	3.3311	.9865	-.0119	2.5570
31,35	442.	4.2708	.7972	.0148	2.5656
32,34	442.	3.1073	.6345	-.0441	2.5566
33	443.	3.2438	.9825	.0840	2.5587
36,40	426.	4.3709	.6729	.0358	2.5669
37,39	424.	3.1132	.5922	.0299	2.5613
38	426.	3.1714	.9096	-.0622	2.5634
14-POST	424.	2.8585	.9852	-.2620	2.6509
19	420.	2.2738	1.1480	.0874	2.6583
21	422.	2.5379	1.0608	-.3354	2.6552
22	422.	3.2014	1.1373	.3155	2.6552
23	421.	2.5416	1.0816	.1966	2.6532
30	420.	2.7524	.9001	.2379	2.6583
31,35	416.	4.2363	.8192	-.0093	2.6623
32,34	417.	3.0665	.7501	-.1208	2.6631
33	419.	3.2124	1.0179	.3340	2.6599
36,40	404.	4.3649	.7158	.0452	2.6522
37,39	401.	3.0876	.6205	-.1334	2.6509
38	406.	3.1379	.8629	.0591	2.6527
1	454.	2.9515	1.3199	.2755	2.6718
2	449.	2.9733	1.0902	.3197	2.6693
3	476.	3.4391	1.2739	.3412	2.6702
4	474.	4.0338	1.1831	.0651	2.6741
5	472.	3.9915	1.1626	.2854	2.6748
6	475.	3.5600	1.1041	.1655	2.6705
10	480.	3.0021	1.3942	.0701	2.6760
14	480.	3.2352	1.4544	.0806	2.6760

Table F.3.10 Item means and correlations with course grade

SPRING 1975 CHEM 131 N = 545

VARIABLE	NUMBER OF RECORDS	MEAN	STANDARD DEVIATION	CORRELATION WITH GRADE	GRADE MEAN
14-PRE	0.	.5000	.5000	.0000	.5000
19	0.	.5000	.5000	.0000	.5000
21	0.	.5000	.5000	.0000	.5000
22	0.	.5000	.5000	.0000	.5000
23	0.	.5000	.5000	.0000	.5000
30	0.	.5000	.5000	.0000	.5000
31,35	0.	.5000	.5000	.0000	.5000
32,34	0.	.5000	.5000	.0000	.5000
33	0.	.5000	.5000	.0000	.5000
36,40	0.	.5000	.5000	.0000	.5000
37,39	0.	.5000	.5000	.0000	.5000
38	0.	.5000	.5000	.0000	.5000
14-POST	226.	3.0310	1.0492	-.3017	2.3319
19	223.	1.9283	1.0564	.0015	2.3318
21	223.	2.9058	1.2035	-.2965	2.3274
22	223.	2.8565	1.0274	.1622	2.3274
23	223.	2.3857	1.0648	.1464	2.3274
30	221.	2.4434	.9382	.2373	2.3281
31,35	219.	4.3503	.8154	-.0313	2.3265
32,34	220.	3.0857	.7628	-.1755	2.3318
33	220.	3.2318	1.0427	.3868	2.3318
36,40	215.	4.3462	.8290	.1185	2.3209
37,39	216.	3.0556	.6955	-.1252	2.3241
38	216.	3.1111	.9797	.2691	2.3241
1	263.	2.8023	1.3506	.3272	2.3156
2	260.	2.8385	1.1389	.3973	2.3115
3	279.	3.2746	1.3150	.2654	2.3172
4	277.	3.7942	1.2593	.1487	2.3159
5	278.	3.6978	1.3094	.2481	2.3165
6	279.	3.4480	1.1438	.2056	2.3172
10	280.	3.2036	1.4084	.1062	2.3125
14	280.	3.0282	1.4699	.0856	2.3125

Table F.3.11 Item means and correlations with course grade

FALL 1975 CHEM 130 N = 1381

VARIABLE	NUMBER OF RECORDS	MEAN	STANDARD DEVIATION	CORRELATION WITH GRADE	GRADE MEAN
14-PRE	1095.	2.7534	.8972	-.0706	2.8187
19	1099.	2.8562	1.1980	.0053	2.8239
21	1103.	2.0816	.8380	-.0382	2.8228
22	1102.	4.0064	.8839	.0190	2.8230
23	1091.	3.5041	1.0750	.0350	2.8236
30	1098.	3.6138	.9371	.0030	2.8201
31,35	1087.	4.4323	.6935	.0350	2.8206
32,34	1085.	3.1830	.6435	-.0162	2.8180
33	1095.	3.3936	.9143	.0581	2.8219
36,40	1071.	4.4643	.6375	.0866	2.8231
37,39	1069.	3.1558	.5673	-.0292	2.8204
38	1075.	3.1712	.8895	.0415	2.8247
14-POST	1138.	2.7250	.9710	-.2697	2.8801
19	1128.	2.4273	1.1910	.1116	2.8848
21	1130.	2.4460	1.0698	-.3124	2.8841
22	1127.	3.5013	1.0772	.2371	2.8838
23	1127.	2.8234	1.1467	.1614	2.8838
30	1121.	2.9955	.9719	.2108	2.8840
31,35	1114.	4.2362	.8119	-.0006	2.8833
32,34	1114.	3.1011	.6975	-.0556	2.8833
33	1118.	3.3417	1.0113	.3197	2.8828
36=40	1096.	4.3702	.7025	.0718	2.8850
37,39	1098.	3.0827	.6288	-.0709	2.8857
38	1100.	3.1309	.9299	.0733	2.8859
1	1109.	3.1488	1.3404	.3123	2.8742
2	1096.	3.1551	1.1047	.3720	2.8750
3	1132.	3.5618	1.2777	.3801	2.8794
4	1132.	4.1087	1.1324	.1663	2.8799
5	1129.	4.0328	1.1103	.2567	2.8822
6	1131.	3.7374	1.0618	.2416	2.8802
10	1135.	3.0529	1.3624	.0162	2.8811
14	1135.	3.0534	1.4885	.0236	2.8811

Table F.3.12 Item means and correlations with course grade

WINTER 1976 CHEM 131 N = 1267

VARIABLE	NUMBER OF RECORDS	MEAN	STANDARD DEVIATION	CORRELATION WITH GRADE	GRADE MEAN
14-PRE	0.	.5000	.5000	.0000	.5000
19	0.	.5000	.5000	.0000	.5000
21	0.	.5000	.5000	.0000	.5000
22	0.	.5000	.5000	.0000	.5000
23	0.	.5000	.5000	.0000	.5000
30	0.	.5000	.5000	.0000	.5000
31,35	0.	.5000	.5000	.0000	.5000
32,34	0.	.5000	.5000	.0000	.5000
33	0.	.5000	.5000	.0000	.5000
36,40	0.	.5000	.5000	.0000	.5000
37,39	0.	.5000	.5000	.0000	.5000
38	0.	.5000	.5000	.0000	.5000
14-POST	771.	3.0597	1.0131	-.2850	2.6459
19	766.	2.0248	1.0879	.0711	2.6482
21	765.	2.8575	1.1757	-.3662	2.6490
22	764.	3.0380	1.0472	.2978	2.6486
23	764.	2.5275	1.0820	.1864	2.6486
30	761.	2.5900	.9356	.3157	2.6491
31,35	748.	4.3048	.7953	-.0155	2.6444
32,34	753.	3.1681	.7132	-.0902	2.6521
33	754.	3.2586	1.0149	.2942	2.6525
36,40	733.	4.4013	.7118	.0206	2.6630
37,39	736.	3.1335	.6395	-.1116	2.6583
38	738.	3.1084	.9608	.0909	2.6599
1	776.	2.9742	1.3582	.3964	2.6405
2	763.	2.9450	1.1168	.4316	2.6370
3	790.	3.3111	1.2955	.4114	2.6462
4	786.	3.7761	1.2624	.2146	2.6559
5	784.	3.9439	1.2014	.2941	2.6492
6	789.	3.5437	1.0855	.2644	2.6470
10	795.	3.1233	1.3570	.0574	2.6484
14	795.	3.0813	1.4507	.0384	2.6484

Table F.3.13 Item means and correlations with course grade

WINTER 1976 CHEM 130 N = 671

VARIABLE	NUMBER OF RECORDS	MEAN	STANDARD DEVIATION	CORRELATION WITH GRADE	GRADE MEAN
14-PRE	467.	3.0000	.9209	-.0214	2.6167
19	471.	2.5648	1.2102	-.0012	2.6062
21	473.	2.3658	.9598	.0730	2.6068
22	470.	3.7319	1.0171	-.0187	2.6053
23	467.	3.2077	1.1210	-.0958	2.6124
30	473.	3.3192	1.0145	.0030	2.6089
31,35	468.	4.3944	.7639	-.0142	2.6015
32,34	470.	3.1155	.6306	.0651	2.6043
33	470.	3.2638	.9240	.0168	2.6053
36,40	462.	4.3878	.7067	.0706	2.6039
37,39	464.	3.1232	.5815	.0525	2.6056
38	464.	3.1659	.8588	.0262	2.6056
14-POST	551.	2.9074	1.0262	-.2601	2.7459
19	545.	2.3046	1.1851	.1421	2.7532
21	550.	2.5782	1.1071	-.3141	2.7482
22	548.	3.2755	1.0282	.2719	2.7482
23	550.	2.6055	1.1516	.2100	2.7482
30	549.	2.8106	.9196	.3123	2.7495
31,35	545.	4.3279	.7806	-.0431	2.7477
32,34	545.	2.9932	.6924	-.0311	2.7514
33	548.	3.3120	1.0106	.2712	2.7491
36,40	535.	4.3698	.6754	.0076	2.7505
37,39	536.	3.0352	.6273	-.0846	2.7509
38	537.	3.1974	.9104	.0947	2.7505
1	539.	3.1169	1.3618	.3302	2.7421
2	531.	3.1017	1.1204	.3945	2.7495
3	557.	3.5063	1.2324	.3435	2.7352
4	557.	4.1903	1.0446	.2115	2.7361
5	555.	4.0144	1.1294	.3356	2.7423
6	554.	3.6354	1.0565	.2905	2.7383
10	558.	3.0305	1.3537	.0281	2.7366
14	558.	3.0896	1.4769	.0636	2.7366

Table F.3.14 Item means and correlations with course grade

SPRING 1976 CHEM 131 N = 575

VARIABLE	NUMBER OF RECORDS	MEAN	STANDARD DEVIATION	CORRELATION WITH GRADE	GRADE MEAN
14-PRE	0.	.5000	.5000	.0000	.5000
19	0.	.5000	.5000	.0000	.5000
21	0.	.5000	.5000	.0000	.5000
22	0.	.5000	.5000	.0000	.5000
23	0.	.5000	.5000	.0000	.5000
30	0.	.5000	.5000	.0000	.5000
31,35	0.	.5000	.5000	.0000	.5000
32,34	0.	.5000	.5000	.0000	.5000
33	0.	.5000	.5000	.0000	.5000
36,40	0.	.5000	.5000	.0000	.5000
37,39	0.	.5000	.5000	.0000	.5000
38	0.	.5000	.5000	.0000	.5000
14-POST	298.	3.1309	1.0553	-.3383	2.4128
19	296.	2.0507	1.1334	.1434	2.4172
21	295.	2.9322	1.1913	-.3766	2.4186
22	295.	2.8678	1.1347	.2824	2.4203
23	295.	2.3932	1.1022	.2399	2.4186
30	294.	2.4660	.9917	.2575	2.4235
31,35	284.	4.2777	.8258	-.0007	2.4296
32,34	287.	3.0786	.6997	-.0013	2.4286
33	289.	3.2111	1.0720	.3171	2.4291
36,40	278.	4.3587	.7616	.0653	2.4263
37,39	276.	3.0228	.6081	-.0695	2.4475
38	282.	3.0887	.9163	.0890	2.4273
1	368.	2.8152	1.2784	.3613	2.4606
2	366.	2.7732	1.0430	.3990	2.4549
3	374.	3.1283	1.3443	.3926	2.4492
4	373.	3.8284	1.2242	.1695	2.4517
5	373.	3.8660	1.2091	.1404	2.4477
6	369.	3.3740	1.1481	.0755	2.4472
10	377.	3.5517	1.2775	-.2567	2.4562
14	377.	3.0875	1.4348	-.3214	2.4562

Table F.4.1 Factor means and correlations with course grade

WINTER 1973 CHEM 130 N = 556

FACTOR	NUMBER OF RECORDS	MEAN	STANDARD DEVIATION	CORRELATION WITH GRADE
1	464.	3.2723	.7501	.0848
2	464.	3.6430	.7970	.0177
3	440.	4.3013	.6530	-.0290
4	442.	3.0485	.5731	.0693
5	445.	3.2921	.7637	.0015
6	474.	3.2679	.7844	.2886
7	450.	3.1556	.9197	.2015
8	320.	4.1804	.7417	.1249
9	347.	2.8444	.6219	-.0903
10	397.	3.4421	.7802	.1956
11	434.	3.1797	.3993	.2440
12	450.	3.8859	.7655	.2150

POSTCOURSE ATTITUDE ITEMS ADJUSTED FOR CHANGE

SURVEY ITEM	NUMBER OF RECORDS	EST PRE	RAW POST	ADJ POST
14	482.	2.8216	2.6203	2.6064
19	474.	2.5570	2.3966	2.4222
21	477.	1.9224	1.9644	2.0963
22	472.	3.7606	3.4873	3.3333
23	473.	3.4884	2.9852	2.8108
30	457.	3.3851	3.0088	2.8864
31,35	408.	4.2241	4.1120	3.9004
32,34	428.	2.9866	2.8184	2.6847
33	455.	3.3912	3.5736	3.5654
36,40	344.	4.3422	4.2924	4.0853
37,39	365.	2.9945	2.8521	2.7320
38	406.	3.2118	3.3202	3.3148

Table F.4.2 Factor means and correlations with course grade

SPRING 1973 CHEM 131 N = 498

FACTOR	NUMBER OF RECORDS	MEAN	STANDARD DEVIATION	CORRELATION WITH GRADE
1	0.	.5000	.5000	.0000
2	0.	.5000	.5000	.0000
3	0.	.5000	.5000	.0000
4	0.	.5000	.5000	.0000
5	0.	.5000	.5000	.0000
6	384.	2.8099	.8859	.3338
7	380.	2.7623	.9163	.2791
8	347.	4.3199	.6467	.1013
9	354.	2.9040	.6459	-.1254
10	360.	3.2542	.8391	.2083
11	400.	3.1550	.4333	.4064
12	418.	3.9569	.9514	.1980

POSTCOURSE ATTITUDE ITEMS ADJUSTED FOR CHANGE

SURVEY ITEM	NUMBER OF RECORDS	EST PRE	RAW POST	ADJ POST
14	386.	3.0026	3.0026	3.0026
19	385.	2.0623	2.0623	2.0623
21	385.	2.6286	2.6286	2.6286
22	386.	3.0440	3.0440	3.0440
23	385.	2.5714	2.5714	2.5714
30	380.	2.6763	2.6763	2.6763
31,35	372.	4.2519	4.2519	4.2519
32,34	375.	2.9230	2.9230	2.9230
33	376.	3.3457	3.3457	3.3457
36,40	350.	4.3959	4.3959	4.3959
37,39	355.	2.8849	2.8849	2.8849
38	361.	3.1496	3.1496	3.1496

Table F.4.3 Factor means and correlations with course grade

FALL 1973 CHEM 130 N = 1221

FACTOR	NUMBER OF RECORDS	MEAN	STANDARD DEVIATION	CORRELATION WITH GRADE
1	1091.	3.4195	.7962	.0215
2	1092.	3.6709	.7846	.0262
3	1059.	4.3916	.6259	.0015
4	1059.	3.0985	.5503	.0473
5	1072.	3.3223	.7852	-.0232
6	987.	3.2985	.7880	.2133
7	990.	3.1862	.8827	.2521
8	909.	4.2956	.6802	.0469
9	939.	2.9744	.6034	-.0280
10	940.	3.3426	.8047	.1624
11	0.	.5000	.5000	.0000
12	0.	.5000	.5000	.0000

POSTCOURSE ATTITUDE ITEMS ADJUSTED FOR CHANGE

SURVEY ITEM	NUMBER OF RECORDS	EST PRE	RAW POST	ADJ POST
14	994.	2.7113	2.5674	2.5756
19	988.	2.9474	2.5860	2.5541
21	994.	1.9507	2.1177	2.2885
22	992.	3.9153	3.5474	3.3615
23	992.	3.5141	2.9244	2.7182
30	992.	3.4647	3.0887	2.9350
31,35	974.	4.3535	4.2297	4.0005
32,34	985.	3.1169	2.9991	2.8956
33	987.	3.4265	3.4671	3.4203
36,40	920.	4.3776	4.3571	4.1752
37,39	943.	3.0821	2.9555	2.8229
38	943.	3.2174	3.2206	3.1900

Table F.4.4 Factor means and correlations with course grade

WINTER 1974 CHEM 131 N = 1110

FACTOR	NUMBER OF RECORDS	MEAN	STANDARD DEVIATION	CORRELATION WITH GRADE
1	0.	.5000	.5000	.0000
2	0.	.5000	.5000	.0000
3	0.	.5000	.5000	.0000
4	0.	.5000	.5000	.0000
5	0.	.5000	.5000	.0000
6	467.	2.8451	.8418	.3368
7	458.	2.7780	.9069	.3244
8	285.	4.3053	.7311	-.0282
9	342.	3.0986	.6391	-.0524
10	383.	3.1658	.8592	.2354
11	694.	3.0660	.4258	.2588
12	717.	3.8224	.9986	.1881

POSTCOURSE ATTITUDE ITEMS ADJUSTED FOR CHANGE

SURVEY ITEM	NUMBER OF RECORDS	EST PRE	RAW POST	ADJ POST
14	475.	2.9011	2.9011	2.9011
19	467.	2.1542	2.1542	2.1542
21	472.	2.7246	2.7246	2.7246
22	470.	3.0702	3.0702	3.0702
23	470.	2.5809	2.5809	2.5809
30	463.	2.6998	2.6998	2.6998
31,35	433.	4.2740	4.2740	4.2740
32,34	418.	3.1135	3.1135	3.1135
33	451.	3.1996	3.1996	3.1996
36,40	310.	4.3272	4.3272	4.3272
37,39	371.	3.0716	3.0716	3.0716
38	396.	3.1263	3.1263	3.1263

Table F.4.5 Factor means and correlations with course grade

WINTER 1974 CHEM 130 N = 780

FACTOR	NUMBER OF RECORDS	MEAN	STANDARD DEVIATION	CORRELATION WITH GRADE
1	62.	3.0215	.8400	.0000
2	89.	3.5281	.8492	-.0417
3	77.	4.4694	.5577	.1933
4	95.	3.0541	.5551	.0492
5	105.	3.3857	.7012	.1819
6	0.	.5000	.5000	.0000
7	0.	.5000	.5000	.0000
8	0.	.5000	.5000	.0000
9	0.	.5000	.5000	.0000
10	0.	.5000	.5000	.0000
11	541.	3.0876	.4462	.1844
12	568.	4.0076	.8611	.1357

POSTCOURSE ATTITUDE ITEMS ADJUSTED FOR CHANGE

SURVEY ITEM	NUMBER OF RECORDS	EST PRE	RAW POST	ADJ POST
14	0.	.5000	.5000	.5000
19	0.	.5000	.5000	.5000
21	0.	.5000	.5000	.5000
22	0.	.5000	.5000	.5000
23	0.	.5000	.5000	.5000
30	0.	.5000	.5000	.5000
31,35	0.	.5000	.5000	.5000
32,34	0.	.5000	.5000	.5000
33	0.	.5000	.5000	.5000
36,40	0.	.5000	.5000	.5000
37,39	0.	.5000	.5000	.5000
38	0.	.5000	.5000	.5000

Table F.4.6 Factor means and correlations with course grade

SPRING 1974 CHEM 131 N = 660

FACTOR	NUMBER OF RECORDS	MEAN	STANDARD DEVIATION	CORRELATION WITH GRADE
1	0.	.5000	.5000	.0000
2	0.	.5000	.5000	.0000
3	0.	.5000	.5000	.0000
4	0.	.5000	.5000	.0000
5	0.	.5000	.5000	.0000
6	397.	2.7767	.8625	.3591
7	395.	2.7215	.8677	.3656
8	372.	4.2197	.7158	.0210
9	371.	2.9600	.6178	-.0202
10	375.	3.2600	.8747	.2890
11	428.	3.0935	.4339	.3255
12	440.	3.9561	.9455	.1724

POSTCOURSE ATTITUDE ITEMS ADJUSTED FOR CHANGE

SURVEY ITEM	NUMBER OF RECORDS	EST PRE	RAW POST	ADJ POST
14	481.	3.0499	3.0499	3.0499
19	398.	2.0402	2.0402	2.0402
21	398.	2.6508	2.6508	2.6508
22	398.	3.0352	3.0352	3.0352
23	398.	2.5327	2.5327	2.5327
30	395.	2.6000	2.6000	2.6000
31,35	392.	4.1385	4.1385	4.1385
32,34	390.	2.9678	2.9678	2.9678
33	393.	3.3384	3.3384	3.3384
36,40	372.	4.2919	4.2919	4.2919
37,39	372.	2.9662	2.9662	2.9662
38	375.	3.1920	3.1920	3.1920

Table F.4.7 Factor means and correlations with course grade

FALL 1974 CHEM 130 N = 1242

FACTOR	NUMBER OF RECORDS	MEAN	STANDARD DEVIATION	CORRELATION WITH GRADE
1	1076.	3.3457	.7967	.0988
2	1100.	3.7188	.8017	.0180
3	1094.	4.4516	.5754	.0005
4	1198.	3.1810	.5532	.0455
5	1188.	3.2798	.7853	-.0021
6	368.	2.9484	.9089	.3154
7	363.	2.8613	.9467	.3294
8	344.	4.2982	.7273	-.0321
9	346.	3.1585	.6250	-.0619
10	349.	3.1461	.9187	.2593
11	546.	2.9875	.4580	.1912
12	580.	3.6264	.8851	.1098

POSTCOURSE ATTITUDE ITEMS ADJUSTED FOR CHANGE

SURVEY ITEM	NUMBER OF RECORDS	EST PRE	RAW POST	ADJ POST
14	379.	2.7098	2.8865	2.9376
19	368.	2.7962	2.3940	2.3760
21	371.	2.0404	2.6550	2.9024
22	372.	3.9677	3.2043	2.9320
23	368.	3.4728	2.6060	2.3602
30	366.	3.5355	2.7432	2.5333
31,35	356.	4.4414	4.2665	4.0276
32,34	357.	3.1577	3.1673	3.0765
33	362.	3.3923	3.2210	3.1464
36,40	347.	4.4681	4.3380	4.1205
37,39	349.	3.1285	3.1465	3.0792
38	349.	3.1547	3.0774	3.0514

Table F.4.8 Factor means and correlations with course grade

WINTER 1975 CHEM 131 N = 1112

FACTOR	NUMBER OF RECORDS	MEAN	STANDARD DEVIATION	CORRELATION WITH GRADE
1	0.	.5000	.5000	.0000
2	0.	.5000	.5000	.0000
3	0.	.5000	.5000	.0000
4	0.	.5000	.5000	.0000
5	0.	.5000	.5000	.0000
6	429.	2.8050	.9223	.2582
7	425.	2.7545	.8824	.2757
8	405.	4.3002	.6880	.0194
9	404.	3.1450	.6046	-.0559
10	406.	3.1232	.9007	.2067
11	513.	2.9973	.7833	.2881
12	537.	3.5525	.8864	.1403

POSTCOURSE ATTITUDE ITEMS ADJUSTED FOR CHANGE

SURVEY ITEM	NUMBER OF RECORDS	EST PRE	RAW POST	ADJ POST
14	437.	2.9794	2.9794	2.9794
19	430.	2.1163	2.1163	2.1163
21	432.	2.7315	2.7315	2.7315
22	433.	3.0739	3.0739	3.0739
23	433.	2.5381	2.5381	2.5381
30	425.	2.6612	2.6612	2.6612
31,35	417.	4.2326	4.2326	4.2326
32,34	417.	3.1816	3.1816	3.1816
33	420.	3.1667	3.1667	3.1667
36,40	406.	4.3540	4.3540	4.3540
37,39	405.	3.1108	3.1108	3.1108
38	406.	3.0764	3.0764	3.0764

Table F.4.9 Factor means and correlations with course grade

WINTER 1975 CHEM 13u N = 634

FACTOR	NUMBER OF RECORDS	MEAN	STANDARD DEVIATION	CORRELATION WITH GRADE
1	447.	3.0157	.8018	.0486
2	438.	3.4254	.8632	-.0091
3	422.	4.3257	.6445	.0245
4	420.	3.1143	.5451	-.0141
5	424.	3.2099	.8324	.0103
6	419.	2.9594	.8346	.2855
7	419.	2.8298	.8804	.2957
8	432.	4.3056	.6835	.0191
9	430.	3.0721	.6206	-.1506
10	436.	3.1786	.8091	.2377
11	449.	3.1207	.8533	.3156
12	469.	3.8621	.8723	.2240

POSTCOURSE ATTITUDE ITEMS ADJUSTED FOR CHANGE

SURVEY ITEM	NUMBER OF RECORDS	EST PRE	RAW POST	ADJ POST
14	424.	2.9363	2.8585	2.8475
19	420.	2.3810	2.2738	2.3261
21	422.	2.3744	2.5379	2.6213
22	422.	3.5403	3.2014	3.0544
23	421.	3.0808	2.5416	2.3905
30	420.	3.1786	2.7524	2.6111
31,35	416.	4.2610	4.2363	4.0643
32,34	417.	3.0925	3.0665	3.0042
33	419.	3.2649	3.2124	3.1852
36,40	404.	4.3621	4.3649	4.1850
37,39	401.	3.0862	3.0876	3.0679
38	406.	3.1626	3.1379	3.1076

Table F.4.10 Factor means and correlations with course grade

SPRING 1975 CHEM 131 N = 545

FACTOR	NUMBER OF RECORDS	MEAN	STANDARD DEVIATION	CORRELATION WITH GRADE
1	0.	.5000	.5000	.0000
2	0.	.5000	.5000	.0000
3	0.	.5000	.5000	.0000
4	0.	.5000	.5000	.0000
5	0.	.5000	.5000	.0000
6	222.	2.6622	.8685	.2610
7	221.	2.5596	.8958	.2037
8	215.	4.3515	.7380	.0497
9	216.	3.0714	.6605	-.1698
10	216.	3.1713	.9309	.3631
11	260.	3.0273	.8573	.3668
12	277.	3.6486	1.0010	.2498

POSTCOURSE ATTITUDE ITEMS ADJUSTED FOR CHANGE

SURVEY ITEM	NUMBER OF RECORDS	EST PRE	RAW POST	ADJ POST
14	226.	3.0310	3.0310	3.0310
19	223.	1.9283	1.9283	1.9283
21	223.	2.9058	2.9058	2.9058
22	223.	2.8565	2.8565	2.8565
23	223.	2.3857	2.3857	2.3857
30	221.	2.4434	2.4434	2.4434
31,35	219.	4.3503	4.3503	4.3503
32,34	220.	3.0857	3.0857	3.0857
33	220.	3.2318	3.2318	3.2318
36,40	215.	4.3462	4.3462	4.3462
37,39	216.	3.0556	3.0556	3.0556
38	216.	3.1111	3.1111	3.1111

Table F.4.11 Factor means and correlations with course grade

FALL 1975 CHEM 130 N = 1381

FACTOR	NUMBER OF RECORDS	MEAN	STANDARD DEVIATION	CORRELATION WITH GRADE
1	1085.	3.3429	.8053	.0439
2	1084.	3.7048	.7692	.0249
3	1062.	4.4498	.5917	.0639
4	1060.	3.1690	.5254	-.0282
5	1073.	3.2829	.7887	.0602
6	1128.	3.0845	.8859	.2752
7	1120.	3.1110	.8786	.2447
8	1090.	4.3072	.6786	.0380
9	1093.	3.0920	.5943	-.0656
10	1097.	3.2397	.8359	.2331
11	1093.	3.1931	.8118	.3370
12	1126.	3.9609	.8388	.2906

POSTCOURSE ATTITUDE ITEMS ADJUSTED FOR CHANGE

SURVEY ITEM	NUMBER OF RECORDS	EST PRE	RAW POST	ADJ POST
14	1138.	2.7566	2.7250	2.7399
19	1128.	2.7447	2.4273	2.4159
21	1130.	2.1788	2.4460	2.6019
22	1127.	3.9104	3.5013	3.2884
23	1127.	3.3860	2.8234	2.6296
30	1121.	3.4844	2.9955	2.8170
31,35	1114.	4.3829	4.2362	4.0486
32,34	1114.	3.1498	3.1011	3.0340
33	1118.	3.3649	3.3417	3.3010
36,40	1096.	4.4505	4.3702	4.1730
37,39	1098.	3.1280	3.0827	3.0249
38	1100.	3.1500	3.1309	3.0848

Table F.4.12 Factor means and correlations with course grade

WINTER 1976 CHEM 131 N = 1267

FACTOR	NUMBER OF RECORDS	MEAN	STANDARD DEVIATION	CORRELATION WITH GRADE
1	0.	.5000	.5000	.0000
2	0.	.5000	.5000	.0000
3	0.	.5000	.5000	.0000
4	0.	.5000	.5000	.0000
5	0.	.5000	.5000	.0000
6	763.	2.7038	.8765	.3040
7	761.	2.7179	.8547	.3168
8	730.	4.3503	.6790	.0021
9	736.	3.1502	.6027	-.1107
10	737.	3.1927	.8648	.2156
11	760.	3.0783	.8451	.3934
12	779.	3.7570	.9113	.3302

POSTCOURSE ATTITUDE ITEMS ADJUSTED FOR CHANGE

SURVEY ITEM	NUMBER OF RECORDS	EST PRE	RAW POST	ADJ POST
14	771.	3.0597	3.0597	3.0597
19	766.	2.0248	2.0248	2.0248
21	765.	2.8575	2.8575	2.8575
22	764.	3.0380	3.0380	3.0380
23	764.	2.5275	2.5275	2.5275
30	761.	2.5900	2.5900	2.5900
31,35	748.	4.3048	4.3048	4.3048
32,34	753.	3.1681	3.1681	3.1681
33	754.	3.2586	3.2586	3.2586
36,40	733.	4.4013	4.4013	4.4013
37,39	736.	3.1335	3.1335	3.1335
38	738.	3.1084	3.1084	3.1084

Table F.4.13 Factor means and correlations with course grade

WINTER 1976 CHEM 130 N = 671

FACTOR	NUMBER OF RECORDS	MEAN	STANDARD DEVIATION	CORRELATION WITH GRADE
1	462.	3.0620	.8452	-.0184
2	464.	3.4102	.8572	-.0408
3	460.	4.3882	.6716	.0268
4	463.	3.1166	.5328	.0602
5	463.	3.2127	.7725	.0245
6	542.	2.9410	.9086	.2883
7	547.	2.8958	.8720	.3095
8	533.	4.3492	.6652	-.0229
9	534.	3.0118	.5884	-.0635
10	537.	3.2561	.8451	.2137
11	528.	3.1742	.8113	.3602
12	552.	3.9481	.8129	.3750

POSTCOURSE ATTITUDE ITEMS ADJUSTED FOR CHANGE

SURVEY ITEM	NUMBER OF RECORDS	EST PRE	RAW POST	ADJ POST
14	551.	3.0200	2.9074	2.8883
19	545.	2.4826	2.3046	2.3315
21	550.	2.4745	2.5782	2.6330
22	548.	3.5493	3.2755	3.1511
23	550.	2.9473	2.6055	2.5225
30	549.	3.1439	2.8106	2.6989
31,35	545.	4.3730	4.3279	4.1816
32,34	545.	3.0938	2.9932	2.9193
33	548.	3.2828	3.3120	3.3012
36,40	535.	4.3816	4.3698	4.2409
37,39	536.	3.1045	3.0352	2.9478
38	537.	3.1695	3.1974	3.1800

Table F.4.14 Factor means and correlations with course grade

SPRING 1976 CHEM 131 N = 575

FACTOR	NUMBER OF RECORDS	MEAN	STANDARD DEVIATION	CORRELATION WITH GRADE
1	0.	.5000	.5000	.0000
2	0.	.5000	.5000	.0000
3	0.	.5000	.5000	.0000
4	0.	.5000	.5000	.0000
5	0.	.5000	.5000	.0000
6	295.	2.6655	.9370	.3412
7	291.	2.5773	.9242	.3056
8	273.	4.3208	.7325	.0379
9	275.	3.0530	.5755	-.0353
10	278.	3.1439	.8696	.2450
11	364.	3.0549	.4080	.2784
12	367.	3.6939	.9347	.1767

POSTCOURSE ATTITUDE ITEMS ADJUSTED FOR CHANGE

SURVEY ITEM	NUMBER OF RECORDS	EST PRE	RAW POST	ADJ POST
14	298.	3.1309	3.1309	3.1309
19	296.	2.0507	2.0507	2.0507
21	295.	2.9322	2.9322	2.9322
22	295.	2.8678	2.8678	2.8678
23	295.	2.3932	2.3932	2.3932
30	294.	2.4660	2.4660	2.4660
31,35	284.	4.2777	4.2777	4.2777
32,34	287.	3.0786	3.0786	3.0786
33	289.	3.2111	3.2111	3.2111
36,40	278.	4.3587	4.3587	4.3587
37,39	276.	3.0228	3.0228	3.0228
38	282.	3.0887	3.0887	3.0887

Appendix G Computer programs

In this appendix are presented listings of computer programs written for the preparation and analysis of test and survey data in these studies. The purposes for which these programs were written are briefly described below; also given are the principal tables in which appear the output from these programs.

ZEUS Program ZEUS calculates the summary statistics for tests and subtests which appear in Tables B.3, D.2, D.3, and D.4.

Six separate programs comprise a package which compiles and analyzes course performance and survey data for each term. These programs are:

PROCESS This program rearranges student test data into an acceptable format for subsequent programs and calculates test means which appear in Appendix E.

PREPARE The student numbers are read backwards from the survey results and the order transposed for the next program.

COMPOZ Program COMPOZ matches the different survey forms to the student records from PROCESS.

FINAL This program calculates summary data for testing and grading, some of which appear in Appendix E and in Table C.3.

AVERAGE Unit exam intercorrelations are calculated by Program AVERAGE; they appear in Table C.3.

MINERVA The survey question and factor means which appear in Tables F.3 and F.4 are prepared by Program MINERVA.

```
      PROGRAM ZEUS(INPUT,OUTPUT,TAPE2=INPUT,TAPE3=OUTPUT,TAPE1)
      COMMON MXT(1)
      DIMENSION MTX(1,1,1)
      EQUIVALENCE(MXT,MTX)
      READ(2,1) NUM
    1 FORMAT(I5)
      NTOT=96*NUM
      CALL SETFL(MXT(NTOT))
      NUM1=2
      NUM2=48
      CALL HERA(MTX,NUM1,NUM2,NUM)
      END


      SUBROUTINE HERA(MTX,NUM1,NUM2,NUM)
      COMMON MXT(1)
      DIMENSION MTX(NUM1,NUM2,NUM),UNIT(101)
      DIMENSION KEY(2,43),IN(6 ),LGT(6),NZ(42),INPT(6),JS(2,500)
      INTEGER PIX(61,50),SUB(6,25),ORD(2,40),DW,PO,EST(4)
      REAL NO,N1,ITEM(2,40),RIT(2,43),DIF(2,43)
      DATA IN/60*0/
      UNIT(101)=UNIT(1)=.0044
      UNIT(100)=UNIT(2)=.0065
      UNIT(99)=UNIT(3)=.0453
      UNIT(98)=UNIT(4)=.0679
      UNIT(97)=UNIT(5)=.0861
      UNIT(96)=UNIT(6)=.1030
      UNIT(95)=UNIT(7)=.1191
      UNIT(94)=UNIT(8)=.1343
      UNIT(93)=UNIT(9)=.1490
      UNIT(92)=UNIT(10)=.1624
      UNIT(91)=UNIT(11)=.1753
      UNIT(90)=UNIT(12)=.1880
      UNIT(89)=UNIT(13)=.2006
      UNIT(88)=UNIT(14)=.2139
      UNIT(87)=UNIT(15)=.2266
      UNIT(86)=UNIT(16)=.2391
      UNIT(85)=UNIT(17)=.2433
      UNIT(84)=UNIT(18)=.2531
      UNIT(83)=UNIT(19)=.2645
      UNIT(82)=UNIT(20)=.2745
      UNIT(81)=UNIT(21)=.2801
      UNIT(80)=UNIT(22)=.2943
      UNIT(79)=UNIT(23)=.2952
      UNIT(78)=UNIT(24)=.3037
      UNIT(77)=UNIT(25)=.3109
      UNIT(76)=UNIT(26)=.3178
      UNIT(75)=UNIT(27)=.3244
      UNIT(74)=UNIT(28)=.3306
      UNIT(73)=UNIT(29)=.3336
      UNIT(72)=UNIT(30)=.3442
      UNIT(71)=UNIT(31)=.3477
      UNIT(70)=UNIT(32)=.3528
      UNIT(69)=UNIT(33)=.3576
      UNIT(68)=UNIT(34)=.3621
      UNIT(67)=UNIT(35)=.3664
      UNIT(66)=UNIT(36)=.3704
      UNIT(65)=UNIT(37)=.3741
      UNIT(64)=UNIT(38)=.3776
      UNIT(63)=UNIT(39)=.3807
```

```
      UNIT(62)=UNIT(42)=.3987
      UNIT(61)=UNIT(41)=.3984
      UNIT(60)=UNIT(42)=.3987
      UNIT(59)=UNIT(43)=.39 8
      UNIT(58)=UNIT(44)=.3988
      UNIT(57)=UNIT(45)=.3984
      UNIT(56)=UNIT(46)=.3988
      UNIT(55)=UNIT(47)=.3987
      UNIT(54)=UNIT(48)=.3978
      UNIT(53)=UNIT(49)=.3984
      UNIT(52)=UNIT(50)=.3988
      UNIT(51)=.3989
      NTOT=96*NUM
      DO 1 LLL=1,NTOT
    1 MXT(LLL)=0
      READ(2,100) MPR,NGZ,MDST,MKR3,MKEY
  100 FORMAT(6I5)
      WRITE(MPR,10)
   10 FORMAT(1H1/////5X*CONTROL CARDS EXACTLY AS READ*//)
   11 FORMAT(1X,6I5)
      WRITE(MPR,11) NUM
      WRITE(MPR,11) MPR,NGZ,MDST,MKR3,MKEY
      READ(2,100) LGTH,NA,NB,MTET,MPIX,MSUB,(LGT(I),I=1,6)
      WRITE(MPR,11) LGTH,NA,NB,MTET,MPIX,MSUB,(LGT(I),I=1,6)
      MOO=MSUB
      READ(2,102) KOA,KOB,KOC
      WRITE(MPR,12) KOA,KOB,KOC
   12 FORMAT(13X,6A10)
      READ(2,101) (KEY(1,I),I=1,LGTH),(KEY(2,J),J=1,LGTH)
      WRITE(MPR,13) ((KEY(I,J),J=1,LGTH),I=1,2)
   13 FORMAT(1X,40I2)
  101 FORMAT(40I2)
      READ(2,101) (ORD(1,I),I=1,LGTH),(ORD(2,J),J=1,LGTH)
      WRITE(MPR,13) ((ORD(I,J),J=1,LGTH),I=1,2)
      DO 171 J=1,MSUB
      J2=LGT(J)
      READ(2,101) (SUB(J,I),I=1,J2)
  171 WRITE(MPR,13) (SUB(J,I),I=1,J2)
      KK=LGTH+1
      KL=LGTH+2
      READ(2,101) (NZ(I),I=1,KL)
      WRITE(MPR,13) (NZ(I),I=1,KL)
      READ(2,102) (INPT(I),I=1,6)
      WRITE(MPR,12) (INPT(I),I=1,6)
  102 FORMAT(12X,6A10)
      NTT=NOPO=0
      NA=NA-NGZ
      NB=NB-NGZ
  103 NOPO=NOPO+1
      READ(2,INPT) (IN(NZ(I)),I=1,KL)
      IF(EOF(2)) 106,104
  104 IF(IN(2)-NA) 103,105
  105 NTT=NTT+1
      JS(1,NTT)=IN(1)
      GO TO 103
  106 CALL SORT(JS,NTT)
      DO 107 I=2,NOPO
  107 BACKSPACE 2
      WRITE(3,14)
   14 FORMAT(1H1)
      NA=NA+NGZ
      NB=NB+NGZ
      IN(50)=1
      IN(51)=NTT/4
      IN(52)=NTT/2
      IN(53)=3*NTT/4
```

```
      IN(54)=NTT
  108 READ(2,INPT) (IN(NZ(I)),I=1,KL)
      IF(EOF(2)) 150,109
  109 IF(NGZ) 131,133
  131 CONTINUE
      DO 132 I=2,KL
  132 IN(I)=NEGZ(IN(I))
  133 IF(IN(2)-NB) 110,112
  110 IF(IN(2)-NA) 108,111
  111 NW=1
      GO TO 113
  112 NW=2
  113 CONTINUE
      DO 117 I=50,53
      K1=IN(I)
  115 IF(IN(1)-JS(1,K1)) 108,118,116
  116 K1=K1+1
      IF(K1-IN(I+1)) 115,115,117
  117 CONTINUE
  118 CONTINUE
      IF(MKEY) 124,123
  123 CONTINUE
      DO 122 I=1,LGTH
      IF(KEY(NW,I)-IN(I+2)) 121,119
  119 MTX(NW,ORD(NW,I),K1)=1
      GO TO 121
  120 MTX(NW,ORD(NW,I),K1)=
  121 MTX(NW,KL,K1)=MTX(NW,KL,K1)+MTX(NW,ORD(NW,I),K1)
  122 CONTINUE
      GO TO 108
  124 CONTINUE


      DO 125 I=1,LGTH
      MTX(NW,ORD(NW,I),K1)=IN(I+2)
  125 MTX(NW,KL,K1)=MTX(NW,KL,K1)+MTX(NW,ORD(NW,I),K1)
      GO TO 108
  128 WRITE(MPR,129) IN(2),IN(1)
  129 FORMAT(5X,*UNMATCHED RECORD   FORM*,I2,* NO.*,I10)
      GO TO 108
  150 IF(NTT) 950,950,151
  151 CONTINUE


  600 WRITE(3,601) NA,NB
  601 FORMAT(1H1//16X,*FORM*,I3,37X,*FORM*,I3)
      DO 602 I=1,LGTH
      KEY(1,ORD(1,I))=I
  602 KEY(2,ORD(2,I))=I
      WRITE(3,603) (I,KEY(1,I),I,KEY(2,I),I=1,LGTH)
  603 FORMAT(///5X,*ITEM*,I3,* ORIGINALLY WAS NO.*,I3,15X,*ITEM*,I3,* OR
     1IGINALLY WAS NO.*,I3/)
  650 CONTINUE


  200 SX=SY=SXY=SX2=SY2=NXY=0
      DO 206 I=1,NTT
      IF(MTX(1,KL,I)) 202,202,201
  201 MTX(1,KK,I)=1
  202 IF(MTX(2,KL,I)) 204,204,203
  203 MTX(2,KK,I)=1
  204 IF(MTX(1,KK,I)*MTX(2,KK,I)-1) 206,205
  205 SX=SX+MTX(1,KL,I)
      SX2=SX2+MTX(1,KL,I)**2
      SY=SY+MTX(2,KL,I)
      SY2=SY2+MTX(2,KL,I)**2
```

```
      SXY=SXY+MTX(1,KL,I)*MTX(2,KL,I)
      NXY=NXY+1
  206 CONTINUE
      LTH=LGTH
      HMA=AMEAN=SX/NXY
      HMB=BMEAN=SY/NXY
      TA=ZZ=(SX2-SX**2/NXY)/(NXY-1.)+1.E-10
      HSA=SIGA=SQRT(ZZ)
      TB=ZZ=(SY2-SY**2/NXY)/(NXY-1.)+1.E-10
      HSB=SIGB=SQRT(ZZ)
      COVAB=(SXY-SX*SY/NXY)/(NXY-1)
      RXY=COVAB/SIGA/SIGB
      PCTA=AMEAN/LGTH*100.
      PCTB=BMEAN/LGTH*100.
      WRITE(3,207) KQA,KQB,KQC
  207 FORMAT(1H1,9(/),20X,3A10/////)
      WRITE(3,208) NA,AMEAN,PCTA,SIGA,TA,NB,BMEAN,PCTB,SIGB,TB,
     1COVAB,NXY,RXY
  208 FORMAT(2(5X,'FORM',I2,' MEAN OF',F6.2,' (',F5.2,')  STD DEV OF',F7
     1.3,'  AND VARIANCE OF',F7.3//), 10X,'THE COVARIANCE BETWEEN THE FO
     1RMS IS',F7.3//)'X,   'STATISTICS ARE BASED ON',I4,' PAIRED RECORDS
     1'///5X,'THE CORRELATION COEFFICIENT IS',F8.4//)
      CALL INTRVL(NXY,RXY)


  300 IF(MPIX) 301,305
  301 MPIX=MPIX-1
      L2T=LGTH+1
      DW=49/LGTH
      DO 304 I=1,KL
  304 KEY(1,I)=KEY(2,I)=0
      DO 306 I=1,NTT
      IF(MTX(1,KK,I)*MTX(2,KK,I)-1) 306,305
  305 MZYX=MTX(1,KL,I)+1
      MXYZ=MTX(2,KL,I)+1
      KEY(1,MZYX)=KEY(1,MZYX)+1
      KEY(2,MXYZ)=KEY(2,MXYZ)+1
  306 CONTINUE
      K1=KEY(1,1)
      DO 303 I=1,2
      DO 303 J=1,L2T
      IF(K1-KEY(I,J)) 302,303,303
  302 K1=KEY(I,J)
  303 CONTINUE
      FTR=60./(K1+8.)
      DO 307 I=1,50
      PIX(61,I)=2R .
      DO 307 J=1,60
  307 PIX(J,I)=2R
      DO 308 I=1,59
      IQ=60-I
  308 IN(IQ)=I/FTR
      DO 317 I=1,LGTH
      K1=KEY(1,I)
      K2=KEY(1,I+1)
      K3=KEY(2,I)
      K4=KEY(2,I+1)
      IQ=DW+1
      DO 311 J=1,IQ
      IP=J-1
      KX=60-FTR*K1-FTR*IP*(K2-K1)/DW
      KY=60-FTR*K3-FTR*IP*(K4-K3)/DW
      KZ=DW*I+IP-DW+1
      IF(KX-KY) 309,310
  309 PIX(KX,KZ)=2R /
      PIX(KY,KZ)=2R =
```

```
      GO TO 311
  310 PIX(KX,KZ)=2R *
  311 CONTINUE
  317 CONTINUE
      URTA=URTB=SKWA=SKWB=MEDA=MEDB=0
      K1=.5*NXY/2.
      K2=K3=0
      DO 322 I=1,LGTH
      IF(K2-K1) 318,318,319
  318 K2=K2+KEY(1,I)
      MEDA=I-1
  319 IF(K3-K1) 320,320,322
  320 K3=K3+KEY(2,I)
      MEDB=I-1
  322 CONTINUE
      DO 323 I=1,L2T
      K4=I*DW
      IQ=K4-DW+1
      PIX(61,IQ)=I+26+I/11*17R?
      URTA=URTA+KEY(1,I)*((I-1-AMEAN)/SIGA)**4
      URTB=URTB+KEY(2,I)*((I-1-BMEAN)/SIGB)**4
      SKWA=SKWA+KEY(1,I)*((I-1-AMEAN)/SIGA)**3
  323 SKWB=SKWB+KEY(2,I)*((I-1)-BMEAN)/SIGB)**3
      URTA=URTA/NXY
      URTB=URTB/NXY
      SKWA=SKWA/NXY
      SKWB=SKWB/NXY
      DW=60/FTP
      WRITE(3,324) NA,MEDA,SKWA,URTA,NB,MEDB,SKWB,URTB
  324 FORMAT(///,2(5X,*FORM*,I2,* HAS A MEDIAN OF*,I3,* SKEWNESS OF*,F7.
     13,* AND KURTOSIS OF*,F7.3//))
      WRITE(3,325) (PIX(61,K),K=1,50),DW,DW,(IN(J),(PIX(J,I),I=1,50),
     1IN(J),J=1,60),(PIX(61,L),L=1,50 )
  325 FORMAT(1H1////4X,50R2/I4,100X,I3,60(/I4,50R2,I3)/4X,50R2)
      WRITE(3,326) NA,NB
  326 FORMAT(///5X,*FORM*,I2,* IS REPRESENTED ON THE GRAPH BY / / / / /*
     1//5X,*FORM*,I2,* IS REPRESENTED ON THE GRAPH BY = = = = =*)
  350 CONTINUE


      IF(MDST) 700,750
  700 MDST=MDST-1
      L2T=LGTH+1
      DW=49/LGTH
      DO 701 I=1,50
      PIX(51,I)=2R .
      DO 701 J=1,59
  701 PIX(I,J)=0
      DO 702 I=1,L2T
      IQ=I*DW-DW+1
  702 PIX(51,IQ)=I+26+I/11*17R?
      DO 704 I=1,NTT
      IF(MTX(1,KK,I)*MTX(2,KK,I)-1) 704,703
  703 IP=(MTX(1,KL,I)+1)*DW-DW+1
      IQ=(MTX(2,KL,I)+1)*DW-DW+1
      PIX(IQ,IP)=PIX(IQ,IP)+1
  704 CONTINUE
      WRITE(3,705) NA,NB
  705 FORMAT(1H1/////5X,*FORM*,I2,* DISTRIBUTED ALONG HORIZONTAL AXIS */
     1/5X,*FORM*,I2,* DISTRIBUTED ALONG VERTICAL AXIS.*//////)
      DO 709 I=1,50
      DO 709 J=1,59
      IF(PIX(I,J)) 707,708
  707 K=PIX(I,J)
      PIX(I,J)=K+27+K/11*17R?
      GO TO 709
```

```
708 PIX(I,J)=2P
709 CONTINUE
    WRITE(3,706) (PIX(51,I),I=1,50),(PIX(51,J),(PIX(J,K),K=1,50),
   1PIX(51,J),J=1,50),(PIX(51,L),L=1,50)
706 FORMAT(4X,50R2,50(/2X,52R2)/4X,50R2)
750 CONTINUE


    IF(MSUB) 401,45
401 SX=SY=SX2=SY2=SXY=NXY=0
    MSUB=MSUB-1
    I=MOO-MSUB
    KM=KL+I
    LGTH=LGT(I)
    DO 404 J=1,NTT
    IF(MTX(1,KK,J)*MTX(2,KK,J)) 402,404
402 CONTINUE
    L2T=LGT(I)
    DO 403 L=1,L2T
    MTX(1,KM,J)=MTX(1,KM,J)+MTX(1,SUB(I,L),J)
403 MTX(2,KM,J)=MTX(2,KM,J)+MTX(2,SUB(I,L),J)
    SX=SX+MTX(1,KM,J)
    SX2=SX2+MTX(1,KM,J)**2
    SY=SY+MTX(2,KM,J)
    SY2=SY2+MTX(2,KM,J)**2
    SXY=SXY+MTX(1,KM,J)*MTX(2,KM,J)
    NXY=NXY+1
404 CONTINUE
    KL=KM
    AMEAN=SX/NXY
    PCTA=AMEAN/LGTH*100.
    BMEAN=SY/NXY
    PCTB=BMEAN/LGTH*100.
    TA=ZZ=(SX2-SX**2/NXY)/(NXY-1)
    SIGA=SQRT(ZZ)
    TB=ZZ=(SY2-SY**2/NXY)/(NXY-1)
    SIGB=SQRT(ZZ)
    COVAB=(SXY-SX*SY/NXY)/(NXY-1)
    RXY=COVAB/SIGA/SIGB
    WRITE(3,14)
    WRITE(3,405) I,(SUB(I,J),J=1,L2T)
405 FORMAT(9(/)5X,*SUBTEST*,I2,* COMPARES ITEMS*,25(I3,*,*))
    WRITE(3,406) NA,AMEAN,PCTA,SIGA,TA,NB,BMEAN,PCTB,SIGB,TB,
   1COVAB,NXY,RXY
406 FORMAT(///2(5X,*FORM*,I2,* MEAN OF*,F6.2,* (*,F5.2,*)  STD DEV OF*
   1,F7.3,  *  AND VARIANCE OF*,F7.3//), 10X,*THE COVARIANCE BETWEEN T
   1HE FORMS IS*,F7.3//1 X,  *STATISTICS ARE BASED ON*,I4,* PAIRED RE
   1CORDS*///5X,*THE CORRELATION COEFFICIENT IS*,F8.4//)
    IF(ABS(RXY)-.2) 410,410,407
407 FTR=15./LGT(I)
    R15=FTR*RXY/(1+(FTR-1)*RXY)
    RPP=RXY*(NXY-1)/(NXY-2)
    RXY=FTR*RPP/(1+(FTR-1)*RPP)
    WRITE(3,408) R15,RPP,RXY
408 FORMAT(5X,*ADJUSTED FOR TEST LENGTH,*/17X,*THE COEFFICIENT IS*,F8.
   14//5X,*ADJUSTED FOR GROUP SIZE,*/17X,*THE COEFFICIENT IS*,F8.4//5X
   1,*ADJUSTED FOR BOTH FACTORS,*/17X,*THE COEFFICIENT IS*,F8.4////
   15X,*CONFIDENCE INTERVALS CALCULATED*/5X,*FOR LENGTH-ADJUSTED COEFF
   1ICIENT*//)
    CALL INTRVL(NXY,R15)
    GO TO 300
410 WRITE(3,411)
411 FORMAT(//5X,*BECAUSE R WAS NOT GREATER THAN 0.20,*/5X,*THE COEFFIC
   1IENT WAS NOT STANDARDIZED*//)
    CALL INTRVL(NXY,RXY)
    GO TO 300
```

```
450 CONTINUE

    LGTH=LTH
    AMEAN=HMA
    BMEAN=HMB
    SIGA=HSA
    SIGB=HSB
    KL=LTH+2
    IF(MTET) 501,485
501 CONTINUE
    DO 502 I=1,LTH
502 KEY(1,I)=KEY(2,I)=ORD(1,I)=ORD(2,I)=JS(1,I)=JS(2,I)=0.
    SXY=1.E-10
    DO 509 I=1,NTT
    IF(MTX(1,KK,I)*MTX(2,KK,I)-1) 509,503
503 CONTINUE
    SXY=SXY+1.
    DO 508 J=1,LTH
    IP=MTX(1,J,I)
    IQ=MTX(2,J,I)
    IF(IP-1) 505,504
504 JS(1,J)=JS(1,J)+MTX(1,KL,I)
505 IF(IQ-1) 507,506
506 JS(2,J)=JS(2,J)+MTX(2,KL,I)
507 KEY(1,J)=KEY(1,J)+IP*(1-IQ)
    KEY(2,J)=KEY(2,J)+IP*IQ
    ORD(1,J)=ORD(1,J)+(1-IP)*(1-IQ)
508 ORD(2,J)=ORD(2,J)+(1-IP)*IQ
509 CONTINUE
    SRPQA=SRPQB=SPQA=SPQB=AVPA=AVPB=0
    WRITE(3,511) KQA,KQB,KQC
511 FORMAT(1H1//14X,*SELECTED ITEM ANALYSIS DATA FOR *,3A10///)
    DO 512 I=1,LTH
    TA=KEY(1,I)+1.E-13
    TB=KEY(2,I)+1.E-13
    TC=ORD(1,I)+1.E-13
    TD=ORD(2,I)+1.E-13
    TXY=TA+TB+TC+TD
    N1=TA+TB+1.E-13
    N0=TXY-N1
    DFA=N1/TXY+1.E-13
    DIF(1,I)=DFA-DFA**2+1.E-13
    AVPA=AVPA+DFA
    DMA=JS(1,I)/N1+1.E-13
    PQ=(DFA+.015)*1*0.
    ZZ=TXY**2-TXY+1.E-10
    RIT(1,I)=(DMA-AMEAN)/SIGA/SQRT(ZZ)*N1/UNIT(PQ)
    IF(RIT(1,I).GT.1.) RIT(1,I)=.99999
    SRPQA=SRPQA+RIT(1,I)**2*DIF(1,I)
    SPQA=SPQA+DIF(1,I)
    N1=TB+TD+1.E-13
    N0=TXY-N1
    DFB=N1/TXY+1.E-13
    DIF(2,I)=DFB-DFB**2+1.E-13
    AVPB=AVPB+DFB
    DMB=JS(2,I)/N1+1.E-13
    PQ=(DFB+.015)*1*0.
    RIT(2,I)=(DMB-AMEAN)/SIGB/SQRT(ZZ)*N1/UNIT(PQ)
    IF(RIT(2,I).GT.1.) RIT(2,I)=.99999
    SRPQB=SRPQB+RIT(2,I)**2*DIF(2,I)
    SPQB=SPQB+DIF(2,I)
    ZZ=DIF(1,I)*DIF(2,I)+1.E-8
    PXY=TB/TXY
    RPHI=(PXY-DFA*DFB)/SQRT(ZZ)
    ZZ=TB*TC/TA/TD+1.E-1
```

```
      DEG=180./(1+SQRT(ZZ))
      DEG=DEG/57.29577951
      RTET=COS(DEG)
      WRITE(3,510) I,NA,DFA,RIT(1,I),RPHI,NB,DFB,RIT(2,I),RTET
  510 FORMAT(5X,'ITEM',I3,'  FORM',I2,'  DIFFICULTY',F6.3,'  ITEM-TEST C
     1ORR',F7.3,'    PHI COEFFICIENT',F7.3//14X,'FORM',I2,'  DIFFICULTY'
     1,F6.3,'  ITEM-TEST CORR',F7.3,'    TETRACHORIC COEF',F6.3////)
  512 CONTINUE
      SA2=SIGA**2
      SB2=SIGB**2
      AVPA=AVPA/LTH
      AVPB=AVPB/LTH
      AVQA=1-AVPA
      AVQB=1-AVPB
      R8A=(SA2-SPQA)/SA2*.5+SQRT(SRPQA/SA2+((SA2-SPQA)/SA2*.5)**2)
      R8B=(SB2-SPQB)/SB2*.5+SQRT(SRPQB/SB2+((SB2-SPQB)/SB2*.5)**2)
      R20A=SXY/(SXY-1)*(SA2-SPQA)/SA2
      R20B=SXY/(SXY-1)*(SB2-SPQB)/SB2
      R21A=SXY/(SXY-1.)*(SA2-LTH*AVPA*AVQA)/SA2
      R21B=SXY/(SXY-1.)*(SB2-LTH*AVPB*AVQB)/SB2
      WRITE(3,513) NA,NB,R8A,R8B,R20A,R20B,R21A,R21B
  513 FORMAT(1H1//25X,'FORM',I2,15X,'FORM',I2///5X,'KR-8',F22.4,F21.4///
     15X,'KR-20',F21.4,F21.4///5X,'KR-21',F21.4,F21.4/)
      WRITE(MPR,514) SIGA,SIGB,SA2,SB2,SPQA,SPQB,SRPQA,SRPQB,AVPA,AVPB,
     1SXY,SXY
  514 FORMAT(///5X,'STD DEV',F19.4,F21.4///5X,'VARIANCE',F18.4,F21.4///
     15X,'SUM(P)(Q)',F17.4,F21.4///5X,'SUM(RBIS)(P)(Q)',F11.4,F21.4///
     15X,'AVERAGE P',F17.4,F21.4///5X,'NO. OF RECORDS',F12.4,F21.4)
  550 CONTINUE


      IF(MKR3) 801,85
  801 CONTINUE
      SRIIPQA=SRIIPQB=1.E-13
      DO 806 L=1,2
      DO 805 I=1,LTH
      ITEM(L,I)=0.
      DO 804 J=1,LTH
      TA=TB=TC=TD=1.E-13
      DO 803 K=1,NTT
      IF(MTX(L,KK,K)-1) 803,802
  802 IP=MTX(L,I,K)
      IQ=MTX(L,J,K)
      TA=TA+IP*(1-IQ)
      TB=TB+IP*IQ
      TC=TC+(1-IP)*(1-IQ)
      TD=TD+(1-IP)*IQ
  803 CONTINUE
      TXY=TA+TB+TC+TD
      ZZ=TB/TA*TC/TD
      ZZ=180./(1+SQRT(ZZ))/57.29577951
      ZZ=COS(ZZ)
      WW=DIF(L,I)*DIF(L,J)+1.E-8
      ZZ=ZZ+(TB-(TA+TB)*(TB+TD)/TXY)/TXY/SQRT(WW)
  804 ITEM(L,I)=ITEM(L,I)+ZZ/2.
  805 ITEM(L,I)=(ITEM(L,I)-1.)/(LTH-1.)
  806 CONTINUE
      ACC=1.-1./LTH
      DO 808 I=1,LTH
      SRIIPQA=SRIIPQA+ITEM(1,I)*DIF(1,I)
  808 SRIIPQB=SRIIPQB+ITEM(2,I)*DIF(2,I)
      R3A=(SA2-SPQA+SRIIPQA)/SA2
      R3B=(SB2-SPQB+SRIIPQB)/SB2
      WRITE(3,810) R3A,R3B,ACC,SRIIPQA,SRIIPQB
  810 FORMAT(///5X,'ESTIMATED KR-3',5X,F7.4,F21.4///5X,
     1'GOODNESS OF FIT',F11.4///5X,'SUM(RII)(P)(Q)',F12.4,F21.4//)
```

```
      WRITE(MPR,812) NA,NB,(I,ITEM(1,I),RIT(1,I),DIF(1,I),ITEM(2,I),
     1RIT(2,I),DIF(2,I),I=1,LTH)
  812 FORMAT(1H1,10(/),5X,*ITEM RELIABILITY STATISTICS*//20X,*FORM *,
     1I2,17X,*FORM *,I2//15X,*RII    RIT    P*Q*,7X,*RII    RIT    P*Q*,
     1 5(/1H+,14X,*R*,6X,*R*,16X,*R*,6X,*R*)/(//5X,*ITEM*,
     1I3,3F7.3,3X,3F7.3))
      DO 814 I=2,LTH
      RIT(1,1)=RIT(1,1)+RIT(1,I)
      RIT(2,1)=RIT(2,1)+RIT(2,I)
      ITEM(1,1)=ITEM(1,1)+ITEM(1,I)
  814 ITEM(2,1)=ITEM(2,1)+ITEM(2,I)
      RIT(1,1)=RIT(1,1)/LTH
      RIT(2,1)=RIT(2,1)/LTH
      ITEM(1,1)=ITEM(1,1)/LTH
      ITEM(2,1)=ITEM(2,1)/LTH
      WRITE(MPR,815) ITEM(1,1),RIT(1,1),ITEM(2,1),RIT(2,1)
  815 FORMAT(//5X,*MEANS  *,2F7.3,10X,2F7.3)
  850 CONTINUE
  950 CONTINUE
      RETURN
      END



      SUBROUTINE SORT (JS,N)
      DIMENSION JS(2,500)
      M=2
      MM=1
      K2=1
  100 M=3-M
      MM=3-MM
      L4=0
      IF(K2-N) 200,850,850
  200 L1=L4+1
      L3=L1+K2
      IF(L3-N) 400,400,750
  400 L2=L3-1
      L4=L2+K2
      IF(L4-N) 450,450,440
  440 L4=N
  450 LL=L1
  900 IF(JS(M,L1)-JS(M,L3)) 950,950,600
  950 JS(MM,LL)=JS(M,L1)
      LL=LL+1
      L1=L1+1
      IF(L1-L2) 900,960,960
  960 CONTINUE
      DO 500 LLL=L3,L4
      JS(MM,LL)=JS(M,LLL)
  500 LL=LL+1
      GO TO 200
  600 JS(MM,LL)=JS(M,L3)
      LL=LL+1
      L3=L3+1
      IF(L3-L4) 900,900,650
  650 CONTINUE
      DO 700 LLL=L1,L2
      JS(MM,LL)=JS(M,LLL)
  700 LL=LL+1
      GO TO 200
  750 CONTINUE
      DO 800 LLL=L1,N
  800 JS(MM,LLL)=JS(M,LLL)
      K2=K2+K2
      GO TO 100
  850 IF(M-1) 870,860
```

```
      RETURN
  870 CONTINUE
      DO 875 LLL=1,N
  875 JS(1,LLL)=JS(2,LLL)
      RETURN
      END


      SUBROUTINE INTRVL(NXY,RXY)
      SEZ=NXY-3.
      SEZ=1/SQRT(SEZ)
      A=1.+RXY
      B=1.-RXY
      Z=A/B
      Z=.5*ALOG(Z)
      A=Z+SEZ*2.
      B=Z-SEZ*2.
      A=A*2.
      B=B*2.
      A=EXP(A)
      B=EXP(B)
      A=(A-1.)/(A+1.)
      B=(B-1.)/(B+1.)
      WRITE(3,10) A,B
   10 FORMAT(/5X,*THE 2SIGMA INTERVAL (95.44%) :*,F6.3,* TO*,F6.3//)
      A=Z+SEZ*3.
      B=Z-SEZ*3.
      A=A*2.
      B=B*2.
      A=EXP(A)
      B=EXP(B)
      A=(A-1.)/(A+1.)
      B=(B-1.)/(B+1.)
      WRITE(3,11) A,B
   11 FORMAT(5X,*THE 3SIGMA INTERVAL (99.74%) :*,F6.3,* TO*,F6.3//)
      RETURN
      END


          IDENT    NEGZ
          ENTRY    NEGZ
AOK       SX6      X1+1
NEGZ      PS
          SA1      X1
          PL       X1,AOK
          NZ       X1,NEGZ
          SX6      0
          EQ       NEGZ
          END
```

```
HAL,RANNER,COM                          POZ      .
MAP,OFF.
ATTACH,PREPARE,PREPARE.
ATTACH,COMPOZ,COMPOZ.
ATTACH,PROCESS,PROCESS.
ATTACH,CHECK,FINAL.
ATTACH,AVERAGE,AVERAGE.
ATTACH,MINERVA,MINERVA.
ATTACH,TAPE10,1DAATA.
COPY,TAPE10,TAPE1.
REWIND,TAPE1.
REWIND,TAPE4.
PROCESS.
RETURN,TAPE1.
REWIND,TAPE4.
COPYCR,TAPE4,TAPE1.
RETURN,TAPE4.
FILE,TAPE4,RT=Z,BT=C,FL=80.
FILE,TAPE5,RT=Z,BT=C,FL=80.
COPYCR,INPUT,TAPE8.
REWIND,TAPE4.
REWIND,TAPE8.
PREPARE.
REWIND,TAPE4.
REWIND,TAPE5.
SORTMRG,O=JUNK.
REWIND,TAPE1.
REWIND,TAPE4.
REWIND,TAPE5.
REWIND,TAPE6.
COMPOZ.
REWIND,TAPE4.
REWIND,TAPE1.
COPYCR,TAPE4,TAPE1.
REWIND,TAPE8.
COPYCR,INPUT,TAPE8.
REWIND,TAPE4.
REWIND,TAPE8.
REWIND,PREPARE.
PREPARE.
REWIND,TAPE4.
REWIND,TAPE5.
SORTMRG,O=JUNK.
REWIND,TAPE1.
REWIND,TAPE4.
REWIND,TAPE5.
REWIND,COMPOZ.
COMPOZ.
RETURN,TAPE7.
REWIND,TAPE4.
REWIND,TAPE1.
COPYCR,TAPE4,TAPE1.
REWIND,TAPE8.
COPYCR,INPUT,TAPE8.
REWIND,TAPE4.
REWIND,TAPE8.
REWIND,PREPARE.
PREPARE.
REWIND,TAPE4.
REWIND,TAPE5.
```

```
SORTMRG,O=JUNK.
REWIND,TAPE1.
REWIND,TAPE4.
REWIND,TAPE5.
REWIND,COMPOZ.
COMPOZ.
RETURN,TAPE1.
REWIND,TAPE4.
.....
CCCCCCCCCC  NEXT SEVENTEEN CONTROL CARDS INCLUDED ONLY FOR
CCCCCCCCCC  THE SIX TFPMS USING NEW-FORM EVALUATION SURVEY
COPYCR,TAPE4,TAPE1.
REWIND,TAPE8.
COPYCR,INPUT,TAPE8.
REWIND,TAPE4.
REWIND,TAPE8.
REWIND,PREPARE.
PREPARE.
REWIND,TAPE4.
REWIND,TAPE5.
SORTMRG,O=JUNK.
REWIND,TAPE1.
REWIND,TAPE4.
REWIND,TAPE5.
REWIND,COMPOZ.
COMPOZ.
RETURN,TAPE1.
REWIND,TAPE4.
.....
CHECK.
REWIND,TAPE1.
AVERAGE.
REWIND,TAPE1.
MINERVA.
APLIB,TT3093,U*.
CCCCCCCCCC
CCCCCCCCCC  END OF SCOPE CONTROL CARDS
CCCCCCCCCC
000000000000000000000*
CCCCCCCCCC
CCCCCCCCCC  BEGINNING OF INPUT DECKS
CCCCCCCCCC

CCCCCCCCCC     FOUR CONTROL CARDS FOR PROCESS
000000000000000000000*

CCCCCCCCCC     PRECOURSE ATTITUDE SURVEY
000000000000000000000*

CCCCCCCCCC     CONTROL CARD FOR PREPARE
000000000000000000000*
SORT
FILE,INPUT=TAPE4(CR),OUTPUT=TAPE5(CR)
FIELD,NBR(1,6,DISPLAY)
KEY,NBR(A,COBOL6)
END
000000000000000000000*

CCCCCCCCCC     CONTROL CARD FOR COMPOZ
000000000000000000000*

CCCCCCCCCC     POSTCOURSE ATTITUDE SURVEY
000000000000000000000*

CCCCCCCCCC     CONTROL CARD FOR PREPARE
000000000000000000000*
```

```
SORT
FILE,INPUT=TAPE4(CR),OUTPUT=TAPE5(CR)
FIELD,NBR(1,6,DISPLAY)
KEY,NBR(A,COBOL6)
END
0000000000000000000000*

CCCCCCCCCC     CONTROL CARD FOR COMPOZ
0000000000000000000000*

CCCCCCCCCC     COURSE EVALUATION SURVEY
0000000000000000000000*

CCCCCCCCCC     CONTROL CARD FOR PREPARE
0000000000000000000000*
SORT
FILE,INPUT=TAPE4(CR),OUTPUT=TAPE5(CR)
FIELD,NBR(1,6,DISPLAY)
KEY,NBR(A,COBOL6)
END
0000000000000000000000*

CCCCCCCCCC     CONTROL CARD FOR COMPOZ
0000000000000000000000*
.........................................
CCCCCCCCCC  NEXT FOUR LEV-ZERO FILES INCLUDED ONLY FOR
CCCCCCCCCC  SIX TERMS USING NEW-FORM EVALUATION SURVEY

CCCCCCCCCC     NEW FORM EVALUATION SURVEY IF APPLICABLE
0000000000000000000000*

CCCCCCCCCC     CONTROL CARD FOR PREPARE
0000000000000000000000*
SORT
FILE,INPUT=TAPE4(CR),OUTPUT=TAPE5(CR)
FIELD,NBR(1,6,DISPLAY)
KEY,NBR(A,COBOL6)
END
0000000000000000000000*

CCCCCCCCCC     CONTROL CARD FOR COMPOZ
0000000000000000000000*
.........................................

CCCCCCCCCC     SIX CONTROL CARDS FOR FINAL CHECK
0000000000000000000000*
CCCCCCCCCC    APLIB ADDFILE CARD
0000000000000000000000*
```

```
      PROGRAM PROCESS(INPUT,OUTPUT,TAPE2=INPUT,TAPE3=OUTPUT,TAPE1,TAPE4)
      DIMENSION MTX(9,2,2),GSC(3,7),IN(9,6),EX(6),AVE(9),AVEX(9)
      INTEGER OUT(73),FNL
      DATA MTX/36*0/,OUT/73*0/
      READ(2,10) NEX,CONV,FCONV,NAME1,NAME2,((GSC(I,J),J=1,7),I=1,3)
   10 FORMAT(I5,2F5.1,4X,2A10/7F5.1/7F5.1/7F5.1)
   98 CONTINUE
      DO 99 I=1,18
   99 OUT(I)=0
      OUT(73)=0
      READ(1,100) OUT(1),FNL
  100 FORMAT(I6,I4)
      IF(EOF(1)) 400,101
  101 XALL=XLST=SUM=SUML=1.E-20
      TNE=TLE=1.E-12
      DO 102 I=1,NEX
      READ(1,103) (IN(I,J),J=1,6)
      IF(EOF(1)) 400,102
  102 CONTINUE
  103 FORMAT(6I5)
      DO 111 I=1,NEX
      DO 105 J=1,6
      Q=NEGZ(IN(I,J))
      IF(Q) 104,106
  104 IN(I,J)=(Q-1)*CONV+.5
CCCCCCCCCC
CCCCCCCCCC                    VALUES IN IN(I,J) ARE STORED AS PERCENTAGES
CCCCCCCCCC                    ONE MUST BE SUBTRACTED FROM 0 BECAUSE OF NEGZ
CCCCCCCCCC                    THE VALUE OF CONV WIL BE 1.0 UNTIL FALL 1975
CCCCCCCCCC
      MTX(I,1,1)=MTX(I,1,1)+IN(I,J)
      MTX(I,1,2)=MTX(I,1,2)+1
      XALL=XALL+IN(I,J)
      OUT(I+5)=OUT(I+5)+1
      TNE=TNE+1.
CCCCCCCCCC                    MTX(I,1,1) IS SUM OF ALL SCORES FOR EXAM (I)
CCCCCCCCCC                    MTX(I,1,2) IS THE NUMBER OF SCORES IN MTX(I,1,1)
CCCCCCCCCC                    NUMBER OF TRIES FOR EXAM I+5 IN OUT(I+5)
CCCCCCCCCC                    TNE IS THE TOTAL NUMBER OF TRIES ON ALL EXAMS
CCCCCCCCCC                    XALL IS THE TOTAL SUM OF ALL EXAMS TAKEN
  105 CONTINUE
  106 IF(J-1) 107,107,109
  107 OUT(I+5)=0
      TLE=TLE+1.
      GO TO 111
CCCCCCCCCC                    SKIPPING AN EXAM PUTS A ZERO IN OUT(I+5) BUT
CCCCCCCCCC                    ADDS ONE TO TLE SO AVERAGE REFLECTS MISSED EXAM
CCCCCCCCCC                    MTX(I,2,2) IS THE NUMBER OF SCORES IN MTX(I,2,1)
CCCCCCCCCC                    MTX(I,2,1) IS SUM OF LAST TRIES FOR EXAM (I)
CCCCCCCCCC
  109 MTX(I,2,1)=MTX(I,2,1)+IN(I,J-1)
      IF(I.EQ.1) OUT(73)=IN(1,J-1)
      MTX(I,2,2)=MTX(I,2,2)+1
      XLST=XLST+IN(I,J-1)
      TLE=TLE+1.
  111 CONTINUE
CCCCCCCCCC                    AT THIS POINT ALL THE SCORES FOR ONE STUDENT HAVE
CCCCCCCCCC                    BEEN READ PROCESSED AND ENTERED IN THE MATRIX.
CCCCCCCCCC                    THE NEXT SECTION CALCULATES MEAN AND STD DEV
      XALL=XALL/TNE+.05
```

```
      XLST=XLST/TLE+.05
      DO 203 I=1,NEX
      IF(OUT(I+5)) 203,203,201
  201 CONTINUE
      JJ=OUT(I+5)
      DO 202 J=1,JJ
      SUM=SUM+(IN(I,J)-XALL)**2
  202 NC=J
      SUML=SUML+(IN(I,NC)-XLST)**2
  203 CONTINUE
      OUT(2)=XALL*10.
      OUT(3)=XLST*10.
      OUT(4)=FNL/FCONV
      OUT(5)=NEX
      SUM=SQRT(SUM/TNE)
      SUML=SQRT(SUML/TLE)
      OUT(17)=SUM*100.
      OUT(18)=SUML*100.
CCCCCCCCCC                    THE LAST-TRIES-AVERAGE WILL BE COMPARED TO THE
CCCCCCCCCC                    GRADING SCALE IN GSC(I) AS  .GT.  SCALE CUTOFF
CCCCCCCCCC                    THE EFFECT OF THE FINAL EXAM WILL BE DETERMINED
CCCCCCCCCC                    USING  .GT. TEN POINT DIFFERENCE FROM SCALE CUTOFF
      DO 305 I=1,7
      IF(OUT(3)-GSC(2,I)*10.) 305,301,301
  301 OUT(15)=OUT(16)=9-I
      IF(OUT(4)-GSC(1,I)) 302,304,304
  302 IF(OUT(4)-GSC(3,I)) 303,307,307
  303 OUT(16)=OUT(15)-1
      GO TO 307
  304 IF(I.EQ.1) GO TO 307
      OUT(16)=OUT(15)+1
      GO TO 307
  305 CONTINUE
      OUT(15)=OUT(16)=1
  307 WRITE(4,399) (OUT(I),I=1,73)
  399 FORMAT(1X,I6,2I4,I3,10I1,2I2,2I4,1X,40I1,14I2,I4)
      GO TO 98
  400 CONTINUE
CCCCCCCCCC                    THE LAST STEP ABOVE WAS TO WRITE THE STUDENTDATA
CCCCCCCCCC                    ON TAPE4 ACCORDING TO STANDARDIZED FORMAT
CCCCCCCCCC                    THE MATRIX OF EXAM SUMS AND FREQUENCIES HAS BEEN
CCCCCCCCCC                    FULLY LOADED AND WILL NEXT BE AVERAGED AND PRINTED
      DO 501 I=1,NEX
      AVEX(I)=AVE(I)=MTX(I,1,2)+1.E-13
  501 AVE(I)=MTX(I,1,1)/AVE(I)
      WRITE(3,502) NAME1,NAME2,(MTX(I,1,1),MTX(I,1,2),AVE(I),I=1,NEX)
  502 FORMAT(1H1,5(/),10X,2A10///10X,*MATRIX FOR ALL TRIES*//10X,*TOTAL
     1    N      MEAN*,9(///4X,I11,* /*,I5,F9.3/))
      DO 503 I=1,NEX
      AVE(I)=MTX(I,2,2)+1.E-13
      AVEX(I)=AVEX(I)/AVE(I)
  503 AVE(I)=MTX(I,2,1)/AVE(I)
      WRITE(3,504) (MTX(I,2,1),MTX(I,2,2),AVE(I),AVEX(I),I=1,NEX)
  504 FORMAT(10(/),10X,*MATRIX FOR LAST TRIES*//10X,*TOTAL      N      MEA
     1N          AVE NO TAKEN*,9(///4X,I11,* /*,I5,F9.3,F19.3/))
      END
          IDENT    NEGZ
          ENTRY    NEGZ
AOK       SX6      X1+1
NEGZ      PS
          SA1      X1
          PL       X1,AOK
          NZ       X1,NEGZ
          SX6      0
          EQ       NEGZ
          END
```

```
      PROGRAM PREPARF(INPUT,OUTPUT,TAPE2=INPUT,TAPE3=OUTPUT,TAPER,TAPE4)
      DIMENSION IN(8),NAME(3)
      READ(2,102) (NAME(I),I=1,3)
102   FORMAT(3A10)
      N=0
103   READ(8,104) N1,N2,N3,N4,N5,N6,(IN(I),I=1,8)
104   FORMAT(6A1,A4,7A10)
      IF(EOF(8)) 300,106
106   WRITE(4,104) N6,N5,N4,N3,N2,N1,(IN(I),I=1,8)
      N=N+1
      GO TO 103
300   WRITE(3,302) (NAME(I),I=1,3),N
302   FORMAT(1H1,10(/),5X,*******  BATCH *,3A10///5X,*******  NUMBER PRE
     )PARED*,I9)
      END
```

```
      PROGRAM COMPOZ(INPUT,OUTPUT,TAPE2=INPUT,TAPE3=OUTPUT,TAPE1,TAPE4,
     1TAPE5,TAPE6,TAPE7)
      INTEGER OUT(73),PROG,PV
      DIMENSION MTX(73),IN(27),NAME(4)
      READ(2,100) (NAME(I),I=1,4),PROG,PV
100   FORMAT(4A10,2I10)
      IF(PROG.LT.1) GO TO 999
      NRD=NFD=NNF=NSUB=0.
101   READ(1,102) (MTX(I),I=1,73)
102   FORMAT(1X,I6,2I4,I3,10I1,2I2,2I4,1X,40I1,14I2,I4)
      IF(EOF(1)) 200,103
103   NRD=NRD+1


198   IF(PROG-2) 199,199,260
199   READ(5,200) N9P,L7,L8,IN(22),L9,L10,L13,L14,IN(2),L15,IN(16),L16,
     1L17,IN(19),L18,L19,L20,L21,IN(4),L22,L23,IN(13),IN(14),IN(15),
     1(IN(LL),LL=8,12),IN(6),IN(7)
200   FORMAT(I6,5I1,5X,I1,1X,I1,1X,8I1,1X,8I1,3X,5I1,29X,I1,4X,I1)
      IF(EOF(5)) 700,201
201   CONTINUE
      DO 202 J=1,22
202   IN(J)=NEGZ(IN(J))
      L7=NEGZ(L7)
      L10=NEGZ(L10)
      IF(L7) 203,207
203   IF(L10) 204,207
204   IN(20)=L7+L10
      IF(IN(20)-10) 208,205,207
205   IN(20)=9
```

```
      GO TO 208
  207 IN(20)=0
  208 L8=NEGZ(L8)
      L9=NEGZ(L9)
      IF(L8) 210,214
  210 IF(L9) 211,214
  211 IN(21)=L8+L9
      IF(IN(21)-10) 215,212,214
  212 IN(21)=9
      GO TO 215
  214 IN(21)=0
  215 L13=NEGZ(L13)
      L14=NEGZ(L14)
      L15=NEGZ(L15)
      IF(L13) 216,220
  216 IF(L14) 217,220
  217 IF(L15) 218,220
  218 IN(1)=7-L13-L14-L15
      GO TO 221
  220 IN(1)=0
  221 L16=NEGZ(L16)
      L19=NEGZ(L19)
      IF(L16) 223,227
  223 IF(L19) 224,227
  224 IN(17)=L16+L19
      IF(IN(17)-10) 228,225,227
  225 IN(17)=9
      GO TO 228
  227 IN(17)=0
  228 L17=NEGZ(L17)
      L18=NEGZ(L18)
      IF(L17) 230,234
  230 IF(L18) 231,234
  231 IN(18)=L17+L18
      IF(IN(18)-10) 235,233,234
  233 IN(18)=9
      GO TO 235
  234 IN(18)=0
  235 L20=NEGZ(L20)
      L21=NEGZ(L21)
      IF(L20) 237,241
  237 IF(L21) 238,241
  238 IN(3)=5-L20-L21
      GO TO 241
  240 IN(3)=0
  241 L22=NEGZ(L22)
      L23=NEGZ(L23)
      IF(L22) 243,246
  243 IF(L23) 244,246
  244 IN(5)=5-L22-L23
      GO TO 247
  246 IN(5)=0
CCCCCCCCCC
CCCCCCCCCC                    IF PROG=1 (PRECOURSE) OR IF PROG=2 (POSTCOURSE)
CCCCCCCCCC                    THEN IN(12) WILL HAVE RESPONSE ADJUSTED
CCCCCCCCCC
  247 IF(PROG-2) 249,252,250
  249 IF(IN(12)-1) 256,256,250
  250 IN(12)=IN(12)-1
      GO TO 256
  252 IF(IN(12)-2) 256,256,250
  256 GO TO 300


  260 READ(5,261) N8S,(IN(I),I=1,14)
  261 FORMAT(I6,3X,14I1)
```

```
      IF(EOF(5)) 700,262
  262 CONTINUE
      DO 263 I=1,14
      IN(I)=NEGZ(IN(I))
      IF(IN(I).GT.5) IN(I)=0
  263 IN(I)=10*IN(I)


  300 IF(NBR.LT.1) GO TO 325
      IF(NBR-MTX(1)) 325,301,325
  301 NFD=NFD+1
      IF(PROG-2) 303,305,309
  303 CONTINUE
      DO 304 I=1,22
  304 MTX(18+I)=IN(I)
      GO TO 320
  305 MTX(19)=IN(1)
      MTX(20)=IN(2)
      MTX(21)=IN(3)
      MTX(22)=IN(4)
      MTX(23)=IN(5)
      DO 306 I=6,22
  306 MTX(35+I)=IN(I)
      GO TO 320
  309 IN(14)=60-IN(14)
      IN(10)=60-IN(10)
      IF(PV-1) 999,314,313
  310 IF(IN(14)) 311,312
  311 IN(14)=IN(14)-3
  312 IF(IN(3)) 313,314
  313 IN(3)=IN(3)-2
  314 MTX(58)=PV
      DO 316 I=1,14
  316 MTX(58+I)=IN(I)
  320 WRITE(4,102) (MTX(I),I=1,73)
      GO TO 101


  325 NSUR=NSUP+1
      DO 326 I=1,72
  326 OUT(I)=0
      IF(PROG-2) 327,329,332
  327 CONTINUE
      DO 328 I=1,22
  328 OUT(18+I)=IN(I)
      GO TO 340
  329 OUT(19)=IN(1)
      OUT(20)=IN(2)
      OUT(21)=IN(3)
      OUT(22)=IN(4)
      OUT(23)=IN(5)
      DO 330 I=6,22
  330 OUT(40+I)=IN(I)
      GO TO 340
  332 IN(14)=60-IN(14)
      IN(10)=60-IN(10)
      IF(PV-1) 999,337,333
  333 IF(IN(14)) 334,335
  334 IN(14)=IN(14)-3
  335 IF(IN(3)) 336,337
  336 IN(3)=IN(3)-2
  337 OUT(58)=PV
      DO 339 I=1,14
  339 OUT(58+I)=IN(I)
  340 WRITE(6,102) (OUT(I),I=1,73)
      BACKSPACE 5
```

```
      READ(5,341) (OUT(I),I=1,8)
  341 FORMAT(8A10)
      WRITE(7,341) (OUT(I),I=1,8)
      GO TO 198


  350 NNF=NNF+1
      WRITE(4,102) (MTX(I),I=1,73)
      READ(1,102) (MTX(I),I=1,73)
      IF(EOF(1)) 850,351
  351 NRD=NRD+1
      GO TO 300


  700 IF(MTX(1)-999999) 701,702,850
  701 NNF=NNF+1
  702 WRITE(4,102) (MTX(I),I=1,73)
      READ(1,102) (MTX(I),I=1,73)
      IF(EOF(1)) 850,703
  703 NRD=NRD+1
      GO TO 701


  800 MTX(1)=999999
      GO TO 198
  850 WRITE(3,851) (NAME(I),I=1,4),NRD,NFD,NNF,NSUR
  851 FORMAT(////////////5X,4A10///5X,*NUMBER READ FROM TAPE*,I5//5X,*MATC
     1HED WITH SURVEYS*,I6//5X,*NO SURVEY FOUND FOR*,I7//5X,*SURVEYS UNM
     1ATCHED*,I9)
      PCT=NRD*.01+1.E-13
      PCT=NFD/PCT
      WRITE(3,852) PCT
  852 FORMAT(//5X,*PERCENTAGE RETURN*,F12.2)
      IF(MTX(1).LT.999999) STOP
      BACKSPACE 4
      ENDFILE 4
      STOP
  999 WRITE(3,1000)
 1000 FORMAT(/* *******/* *******/* *******/* *******/* *******/6X,*PROGRAM I
     1NPUT PARAMETER(S) ABSENT - JOB ABORTED*)
      END


          IDENT     NEGZ
          ENTRY     NEGZ
AOK       SX6       X1+1
NEGZ      PS
          SA1       X1
          PL        X1,AOK
          NZ        X1,NEGZ
          SX6       0
          EQ        NEGZ
          END
```

```
      PROGRAM FINAL(INPUT,OUTPUT,TAPE2=INPUT,TAPE3=OUTPUT,TAPE1,TAPE4)
      DIMENSION IN(73),NAME(8),FREQ(9,6),GSC(3,7)
      INTEGER GRADE(4,9),EGS(9,67),GD
      DATA FREQ/54*0/,GRADE/36*0/,EGS/603*0/
      READ(2,501) (NAME(I),I=1,8),NEX
  501 FORMAT(3A10,I1)
      NTOT=NREAL=NFNL=FNL=RTOT=RLST=RBF=RAF=0.
  101 READ(4,502) (IN(I),I=1,73)
  502 FORMAT(1X,I6,2I4,I3,1 I),2I2,2I4,1X,40I1,14I2,I4)
      IF(EOF(4)) 200,102
  102 NTOT=NTOT+1
      NUM=0
      DO 104 I=1,NEX
      IF(IN(5+I)) 104,103
  103 NUM=NUM+I
  104 CONTINUE
      IF(IN(4)) 106,105
  105 NUM=NUM+NEX
      GO TO 107
  106 FNL=FNL+IN(4)
      NFNL=NFNL+1
  107 IF(NUM-2*NEX) 108,119,119
  108 NREAL=NREAL+1
      FG=IN(15)
      IF(FG.LT.1.9) FG=0.
      RBF=RBF+FG/2.
      FG=IN(16)
      IF(FG.LT.1.9) FG=0.
      RAF=RAF+FG/2.
      RTOT=RTOT+IN(2)
      RLST=RLST+IN(3)
      DO 110 I=1,NEX
      NO=IN(5+I)
      IF(NO) 109,110
  109 FREQ(I,NO)=FREQ(I,NO)+1
  110 CONTINUE
      IQ=9-IN(15)
      IF(IN(15)-IN(16)) 111,112,113
  111 GRADE(4,IQ)=GRADE(4,IQ)+1
      GRADE(4,9)=GRADE(4,9)+1
      GO TO 120
  112 GRADE(3,IQ)=GRADE(3,IQ)+1
      GRADE(3,9)=GRADE(3,9)+1
      GO TO 120
  113 GRADE(2,IQ)=GRADE(2,IQ)+1
      GRADE(2,9)=GRADE(2,9)+1
      GO TO 120
  119 IN(15)=IN(16)=0
      IQ=9
      IF(IN(2)) 120,121
  120 GRADE(1,IQ)=GRADE(1,IQ)+1
      WRITE(1,502) (IN(I),I=1,73)
      GD=(IN(2)+5)/10
      GD=101-GD
      IF(GD.GT.66) GD=66
      IQ=9-IN(16)
      IF(IQ.GT.8) GO TO 10
      EGS(IQ,GD)=EGS(IQ,GD)+1
      EGS(IQ,67)=EGS(IQ,67)+1
      EGS(9,GD)=EGS(9,GD)+1
      EGS(9,67)=EGS(9,67)+1
```

```
      GO TO 101
  200 FNL=FNL/NFNL
      RBF=RBF/NREAL
      RAF=RAF/NREAL
      RTOT=RTOT/NREAL/10.
      RLST=RLST/NREAL/10.
      P2=GRADE(2,9)*100./NREAL
      P3=GRADE(3,9)*100./NREAL
      P4=GRADE(4,9)*100./NREAL
      DO 202 I=1,NEX
      DO 202 J=1,6
  202 FREQ(I,J)=FREQ(I,J)/NREAL*100.
      WRITE(3,503) (NAME(I),I=1,3),NTOT,NREAL,RTOT,RBF,RLST,RAF,NFNL,FNL
  503 FORMAT(1H1/////5X,3A1 ///5X,*NUMBER OF STUDENTS ON INPUT*,I9//
     15X,*NUMBER OF COURSE COMPLETIONS*,I8///5X,*NET MEAN ON ALL TRIES*,
     1F10.3,20X,*PRE-F  G:*,F7.4//5X,*NET MEAN ON LAST TRIES*,F10.3,20X,
     1*POST-F G:*,F7.4////5X,*NUMBER WHO TOOK FINAL EXAM*,
     1I10//5X,*MEAN FINAL EXAM SCORE*,F10.3)
      READ(2,506) (NAME(I),I=1,8),((GSC(J,K),K=1,7),J=1,3)
  506 FORMAT(8A10/7F5.1/7F5.1/7F5.1)
      WRITE(3,504)((GRADE(I,J),I=1,4),GSC(2,J),GSC(1,J),GSC(3,J),J=1,7),
     1((GRADE(K,L),K=1,4),L=8,9),P2,P3,P4
  504 FORMAT(/////5X,*DISTRIBUTION*,8X,*NUMBER OF GRADES:*,6X,*SCALE*,
     16X,*RAISE  KEEP*/5X,*BEFORE FINAL    LOWERD   UNCHGD   RAISED   CU
     1TOFF      POINT  POINT*//7X,*4.0*,I6,3I9,F10.1,3X,2F7.0//
     17X,*3.5*,I6,3I9,F10.1,3X,2F7.0//7X,*3.0*,I6,3I9,F10.1,3X,2F7.0//
     17X,*2.5*,I6,3I9,F10.1,3X,2F7.0//7X,*2.0*,I6,3I9,F10.1,3X,2F7.0//
     17X,*1.5*,I6,3I9,F10.1,3X,2F7.0//7X,*1.0*,I6,3I9,F10.1,3X,2F7.0//
     17X,*0.0*,I6,3I9//5X,*OTHER*,I6,3I9//19X,3F9.2)
      WRITE(3,505) (K,K=1,6),(I,(FREQ(I,J),J=1,6),I=1,NEX)
  505 FORMAT(/////10X,*DISTRIBUTION OF EXAM TAKING FREQUENCIES*//
     15X,*TAKES*,6I9/7X,*EXAM*//9(6X,I1,6X,6F9.3//))
      WRITE(3,507) (NAME(I),I=1,8),(GSC(2,J),J=1,7),IQ
  507 FORMAT(1H1/////7X,8A1 ///13X,*4.0*,5X,*3.5*,5X,*3.0*,5X,*2.5*,
     15X,*2.0*,5X,*1.5*,5X,*1.0*,5X,*0.0*//8X,7F8.1//11QX,I20)
      DO 210 M=1,66
      J=101-M
  210 WRITE(3,508) J,(EGS(L,M),L=1,9)
  508 FORMAT(3X,I5,9I8)
      WRITE(3,509) (EGS(I,67),I=1,9)
  509 FORMAT(1H0,7X,9I8)
      REWIND 1
      DO 212 I=1,9
      DO 212 J=1,67
  212 EGS(I,J)=0
      READ(2,506) (NAME(I),I=1,8)
  215 READ(1,510) (IN(I),I=1,7)
  510 FORMAT(A7,2I4,I3,A10,2I2)
      IF(EOF(1)) 220,216
  216 IF(IN(6)) 217,215
  217 GD=(IN(3)+5)/10
      GD=101-GD
      IF(GD.GT.66) GD=66
      IQ=9-IN(7)
      IF(IQ.GT.8) GO TO 215
      EGS(IQ,GD)=EGS(IQ,GD)+1
      EGS(IQ,67)=EGS(IQ,67)+1
      EGS(9,GD)=EGS(9,GD)+1
      EGS(9,67)=EGS(9,67)+1
      GO TO 215
  220 WRITE(3,507) (NAME(I),I=1,8),(GSC(2,J),J=1,7),IQ
      DO 221 M=1,66
      J=101-M
  221 WRITE(3,508) J,(EGS(L,M),L=1,9)
      WRITE(3,509) (EGS(I,67),I=1,9)
      END
```

```
      PROGRAM AVERAGE(INPUT,OUTPUT,TAPE2=INPUT,TAPE3=OUTPUT,TAPE1)
      REAL NUM(60,2),LST,VAL(60,2)
      DATA VAL/120*0./,NUM/120*1.E-10/
      ATOT=BTOT=TOT=SUMA=SUMB=AVEA=AVEB=1.E-10
  101 READ(1,501) ALL,LST,NGD,SIGA,SIGB
  501 FORMAT(7X,2F4.1,15X,I2,2F4.2)
      IF(EOF(1)) 110,102
  102 IF(NGD.EQ.0) GO TO 101
      TOT=TOT+1.
      ATOT=ATOT+ALL
      IF(ALL-41.) 103,104,104
  103 ALL=41.
  104 ALL=ALL-39.5
      NA=ALL
      SUMA=SUMA+SIGA**2
      VAL(NA,1)=VAL(NA,1)+SIGA
      NUM(NA,1)=NUM(NA,1)+1.
      BTOT=BTOT+LST
      IF(LST.LE.41.) LST=41.
      LST=LST-39.5
      NB=LST
      SUMB=SUMB+SIGB**2
      VAL(NB,2)=VAL(NB,2)+SIGB
      NUM(NB,2)=NUM(NB,2)+1.
      GO TO 101
  110 WRITE(3,502)
  502 FORMAT(1H1,20(/),18X,*ALL TRIES*,6X,*LAST TRIES*)
      DO 112 I=1,60
      J=61-I
      VAL(J,1)=VAL(J,1)/NUM(J,1)
      VAL(J,2)=VAL(J,2)/NUM(J,2)
      K=40+J
  112 WRITE(3,503) K,VAL(J,1),NUM(J,1),VAL(J,2),NUM(J,2)
  503 FORMAT(/7X,I4,F10.2,2X,F4.0,F10.2,2X,F4.0)
      REWIND 1
      AVEA=ATOT/TOT
      AVEB=BTOT/TOT
      ATOT=BTOT=1.E-10
  119 READ(1,501) ALL,LST,NGD
      IF(EOF(1)) 130,120
  120 IF(NGD.EQ.0) GO TO 119
      ATOT=ATOT+(ALL-AVEA)**2
      BTOT=BTOT+(LST-AVEB)**2
      GO TO 119
  130 RA=1-SUMA/(SUMA+ATOT)
      RB=1-SUMB/(SUMB+BTOT)
      SUMA=SUMA/TOT
      SUMB=SUMB/TOT
      ATOT=ATOT/TOT
      BTOT=BTOT/TOT
      WRITE(3,504) TOT,TOT+AVEA,AVEB,SUMA,SUMB,ATOT,BTOT,RA,RB
  504 FORMAT(//7X,*TOTAL N=*,F12.0,F15.0//4X,*MEAN SCORE=*,F9.4,F15.4
     1//8X,*WITHIN=*,F9.4,F15.4//7X,*BETWEEN=*,F9.4,F15.4
     1//5X,*INTERCORR=*,F9.4,F15.4)
      SUMA=SQRT(SUMA)
      SUMB=SQRT(SUMB)
      ATOT=SQRT(ATOT)
      BTOT=SQRT(BTOT)
      WRITE(3,505) SUMA,SUMB,ATOT,BTOT
  505 FORMAT(///6X,*SEM WTHN=*,F9.4,F15.4//5X,*SEM BETWN=*,F9.4,F15.4)
      END
```

```
      PROGRAM MINERVA(INPUT,OUTPUT,TAPE2=INPUT,TAPE3=OUTPUT,TAPE1)
      DIMENSION NAME(3),COR(6,12),CORR(6,32),CHG(12,3),FAK(12)
      REAL IN(32)
CCCCCCCCCC
CCCCCCCCCC                    TITLE ON INPUT IN COLUMNS 1 TO 36
CCCCCCCCCC
      READ(2,500) (NAME(I),I=1,3)
  500 FORMAT(3A12)
  100 READ(1,502) GRD,(IN(I),I=1,32)
  502 FORMAT(30X,F2.0,14X,2F1.0,1X,3F1.0,4X,9F1.0,1X,3F1.0,4X,7F1.0,
     11X,6F2.1,6X,F2.1,6X,F2.1)
      IF(EOF(1)) 200,101
  101 GRD=GRD/2.
CCCCCCCCCC                    COMPILATION OF FACTORS FROM SURVEYS
CCCCCCCCCC
CCCCCCCCCC
      FAK(1)=(9.-IN(1)+IN(2)-IN(3))/3.+1.
      IF(IN(1)*IN(2)*IN(3).EQ.0) FAK(1)=0.
      FAK(2)=(IN(4)+IN(5)+IN(6)-3.)/3.+1.
      IF(IN(4)*IN(5)*IN(6).EQ.0) FAK(2)=0.
      FAK(3)=(IN(7)+IN(10)-4.)*2./7.+1.
      IF(IN(7).NE.0) IN(7)=(IN(7)-2.)*4./7.+1.
      IF(IN(10).NE.0) IN(10)=(IN(10)-2.)*4./7.+1.
      IF(IN(7)*IN(10).EQ.0) FAK(3)=0.
      FAK(4)=(IN(8)+IN(11)-4.)*2./7.+1.
      IF(IN(8).NE.0) IN(8)=(IN(8)-2.)*4./7.+1.
      IF(IN(11).NE.0) IN(11)=(IN(11)-2.)*4./7.+1
      IF(IN(8)*IN(11).EQ.0) FAK(4)=0.
      FAK(5)=(IN(9)+IN(12)-2.)/2.+1.
      IF(IN(9)*IN(12).EQ.0) FAK(5)=0.
      FAK(6)=(9.-IN(13)+IN(14)-IN(15))/3.+1.
      IF(IN(13)*IN(14)*IN(15).EQ.0) FAK(6)=0.
      FAK(7)=(IN(16)+IN(17)+IN(18)-3.)/3.+1.
      IF(IN(16)*IN(17)*IN(18).EQ.0) FAK(7)=0.
      FAK(8)=(IN(19)+IN(22)-4.)*2./7.+1.
      IF(IN(19).NE.0) IN(19)=(IN(19)-2.)*4./7.+1.
      IF(IN(22).NE.0) IN(22)=(IN(22)-2.)*4./7.+1.
      IF(IN(19)*IN(22).EQ.0) FAK(8)=0.
      FAK(9)=(IN(20)+IN(23)-4.)*2./7.+1.
      IF(IN(20).NE.0) IN(20)=(IN(20)-2.)*4./7.+1.
      IF(IN(23).NE.0) IN(23)=(IN(23)-2.)*4./7.+1.
      IF(IN(20)*IN(23).EQ.0) FAK(9)=0.
      FAK(10)=(IN(21)+IN(24)-2.)/2.+1.
      IF(IN(21)*IN(24).EQ.0) FAK(10)=0.
      FAK(11)=(IN(25)+IN(26)+IN(27)+IN(31)+IN(32)-5.)/5.+1.
      IF(IN(25)*IN(26)*IN(27)*IN(31)*IN(32).EQ.0) FAK(11)=0.
      FAK(12)=(IN(28)+IN(29)+IN(30)-3.)/3.+1.
      IF(IN(28)*IN(29)*IN(30).EQ.0) FAK(12)=0.
CCCCCCCCCC
CCCCCCCCCC
CCCCCCCCCC                    CORRELATION OF FACTORS WITH GRADE
      DO 120 I=1,12
      IF(FAK(I).EQ.0) GO TO 120
      COR(1,I)=COR(1,I)+GRD
      COR(2,I)=COR(2,I)+GRD**2
      COR(3,I)=COR(3,I)+FAK(I)
      COR(4,I)=COR(4,I)+FAK(I)**2
      COR(5,I)=COR(5,I)+GRD*FAK(I)
      COR(6,I)=COR(6,I)+1.
  120 CONTINUE
```

```
CCCCCCCCCC
CCCCCCCCCC                    CORRELATION OF SURVEY ITEMS WITH GRD
CCCCCCCCCC
      DO 140 I=1,32
      IF(IN(I).EQ.0) GO TO 140
      CORR(1,I)=CORR(1,I)+GRD
      CORR(2,I)=CORR(2,I)+GRD**2
      CORR(3,I)=CORR(3,I)+IN(I)
      CORR(4,I)=CORR(4,I)+IN(I)**2
      CORR(5,I)=CORR(5,I)+IN(I)*GRD
      CORR(6,I)=CORR(6,I)+1.
  140 CONTINUE
CCCCCCCCCC
CCCCCCCCCC                    CALCULATION OF PRE/POST CHANGE BONUSES
      DO 170 J=13,24
      I=J-12
      A=IN(I)
      B=IN(J)
      IF(B.EQ.0) GO TO 170
      IF(A.EQ.0) A=B
      IF(B-3.) 158,155,151
  151 IF(A-3.) 152,153,154
  152 CHG(I,1)=CHG(I,1)+5+(B-A-1.)/4.
      GO TO 165
  153 CHG(I,1)=CHG(I,1)+3.5+(B-A)*A/B
      GO TO 165
  154 CHG(I,1)=CHG(I,1)+B-(B-A+1.E-10)**2/2.
      GO TO 165
  155 IF(A-3.) 156,154,157
  156 CHG(I,1)=CHG(I,1)+1.25+B-A/2.
      GO TO 165
  157 CHG(I,1)=CHG(I,1)+1.75+B-A/2.
      GO TO 165
  158 IF(A-3.) 159,160,161
  159 CHG(I,1)=CHG(I,1)+B+(B-A+1.E-10)**2/2.
      GO TO 165
  160 CHG(I,1)=CHG(I,1)+1.5+B/2.-A/4.
      GO TO 165
  161 CHG(I,1)=CHG(I,1)+1.+(B-A+1.)/4.
  165 CHG(I,2)=CHG(I,2)+1.
      CHG(I,3)=CHG(I,3)+A
  170 CONTINUE
      GO TO 180
CCCCCCCCCC
CCCCCCCCCC
CCCCCCCCCC
  200 CONTINUE
      WRITE(3,505) (NAME(I),I=1,3)
  505 FORMAT(1H1////19X,3A1 /////9X,*VARIABLE NUMBER OF     MEAN      STAN
     1DARD  CORRELATION  GRADE*/
     119X,*RECORDS*,19X,*DEVIATION  WITH GRADE    MEAN*)
      DO 240 I=1,32
      CORR(6,I)=CORR(6,I)+1.E-10
      R=CORR(2,I)-CORR(1,I)**2/CORR(6,I)
      R=R*(CORR(4,I)-CORR(3,I)**2/CORR(6,I))
      R=R+1.E-10
      R=(CORR(5,I)-CORR(1,I)*CORR(3,I)/CORR(6,I))/SQRT(R)
      CORR(1,I)=CORR(1,I)/CORR(6,I)
      CORR(3,I)=CORR(3,I)/CORR(6,I)
      SIG=CORR(4,I)/CORR(6,I)-CORR(3,I)**2
      SIG=SQRT(SIG)
  240 WRITE(3,506) I,CORR(1,I),CORR(3,I),SIG,R,CORR(1,I)
  506 FORMAT(/19X,I4,F11.3,F11.4,F11.4,F11.4,F11.4)
CCCCCCCCCC
CCCCCCCCCC
      WRITE(3,501) (NAME(I),I=1,3)
```

```
  501 FORMAT(1H1///1X,3A) ////
     110X,*FACTOR  NUMBER OF     MEAN      STANDARD  CORRELATION*/
     119X,*RECORDS              DEVIATION  WITH GRADE*)
      DO 220 I=1,12
      COR(6,I)=COR(6,I)+1.E-10
      AVE=COR(3,I)/COR(6,I)
      SIG=COR(4,I)/COR(6,I)-AVE**2
      SIG=SQRT(SIG)
      R=(COR(2,I)-COR(1,I)**2/COR(6,I))*(COR(4,I)-COR(3,I)**2/COR(6,I))
      R=R+1.E-10
      R=(COR(5,I)-COR(1,I)*COR(3,I)/COR(6,I))/SQRT(R)
  220 WRITE(3,504) I,COR(6,I),AVE,SIG,R
  504 FORMAT(/10X,I4,F11.0,F11.4,F11.4,F11.4)
CCCCCCCCCC
CCCCCCCCCC
      WRITE(3,503)
  503 FORMAT(////10X,*POSTCOURSE ATTITUDE ITEMS ADJUSTED FOR CHANGE*////
     110X,*SURVEY  NUMBER OF      EST         RAW         ADJ*/
     111X,*ITEM    RECORDS       PRE        POST        POST*)
      DO 230 I=1,12
      J=12+I
      CHG(I,2)=CHG(I,2)+1.E-10
      CHG(I,1)=CHG(I,1)/CHG(I,2)
      CHG(I,3)=CHG(I,3)/CHG(I,2)
  230 WRITE(3,504) I,CHG(I,2),CHG(I,3),COPR(3,J),CHG(I,1)
      END
```